Socially Responsible and Factual Reasoning for Equitable AI Systems

Saadia Gabriel

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Yejin Choi, Co-Chair
Franziska Roesner, Co-Chair
Noah Smith

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Socially Responsible and Factual Reasoning
for Equitable AI Systems

Saadia Gabriel

Co-chairs of the Supervisory Committee:

Brett Helsel Professor Yejin Choi
Computer Science and Engineering

Associate Professor Franziska Roesner
Computer Science and Engineering

Through natural language communication, writers have enormous persuasive power over readers. This can have broad-reaching positive societal impact like in the case of social movements (e.g. the Black Lives Matter movement and protests against anti-Asian hate), however there are severe negative ramifications when communication is used with malintent (e.g. to directly inflict harm through hate speech or mislead). The ability to *read between the lines of what is explicitly stated* and *adapt to dynamic social contexts* is critical to detecting false or harmful text. However, existing deep learning approaches still have limited generalization and commonsense reasoning capabilities.

To expand machine reasoning capabilities, we propose theoretical formalisms to measure intent, factuality and social bias of language. We first introduce reaction frames, which allow us to distill knowledge of cognitive and physical effects on readers like implied actions (e.g. given the false statement "*Water boiled with garlic cures coronavirus*," we can infer that the writer is compelling an audience to "*drink garlic water*"). We find that while neural misinformation detection classifiers are highly capable of distinguishing between truthful and false content, these models are challenged by commonsense implications derived us-

ing our neuro-symbolic approach. We discuss how a major bottleneck comes from the inability of neural models to correctly interpret meaning, particularly when this pertains to plausibility of claims. We conduct a meta-evaluation to test efficacy of factuality metrics, and expose that the evaluation used for generation is ill-suited to benchmarking progress in learning factuality. This study pinpoints specific failure cases of metrics and underlying models, outlining future directions for factuality evaluation.

Finally we show how, despite their limitations, large pretrained language models like GPT-3 can be used to mitigate dataset bias in existing hate speech corpora. We use adversarial generation approaches to better align classifiers with human interpretation of toxicity and mitigate potentially harmful vulnerabilities in classifiers. As future work, we discuss the need for a proactive, community-driven approach to reduce online harms.

Acknowledgements

First and foremost I want to thank my wonderful advisors Yejin Choi and Franziska Roesner, for their invaluable advice, support, wisdom, many eye-opening research meetings and seemingly inexhaustible enthusiasm. Their confidence in me kept me going over the years. I also want to thank Noah Smith for being a great conscientious collaborator and mentor, as well as serving on my doctoral committee. Thank you to Shane Steinert-Threlkeld for agreeing to be the GSR on the committee! Tremendous thanks to many other collaborators, mentors and friends who helped bring this document into existence. I particularly want to note Sofia Serrano and Nicasia Beebe-Wang (my officemates for life), many other wonderful friends like Erin Wilson, Matt Johnson, Esther Jang, Peter West, Willie Agnew, Tal August, Lucy Lin, Elizabeth Clark and Brian Hou for all the fun and food, members of UW BGSA for existing and being honest, members of the xLab, AI2 Mosaic, UW Security & Privacy, and Ark groups for many hours of feedback over the years, Maarten Sap, Max Forbes and Rowan Zellers for years of great conversations, Hannah Rashkin for sisterly advice and all the tea, Yonatan Bisk for being there since literally day 1, Swabha Swayamdipta and Suchin Gururangan for being awesome, and Leilani Battle for brilliant book recommendations and boardgames. Thank you to Les Sessoms for all the encouraging Slack messages. Thank you to Kate Saenko and Ed Lazowska for pushing me speak up. Thank you to Hanna Hajishirzi, Luke Zettlemoyer, Vered Shwartz, Hamid Palangi, Ajay Divakaran, Jan Harrison and Asli Celikyilmaz for their mentorship as I was getting settled in graduate school. Finally, thank you to my dog Raven for not forgetting me after all the time I spent on this thesis. My research was generously supported by the Washington Research Foundation and Ron Howell, as well as funding from NSF, DARPA, Google and Microsoft Research.

DEDICATION

To Ella and Satyananda Gabriel

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ultimate correct theory will have to provide for the ways in which the interpreter of a text acts creatively to build a detailed model of the text. - From Fillmore [1976]

Language development is one of the most powerful abilities we have as humans. It can be an unifying force across geographical and cultural divisions. However, there is a risk of language being weaponized to support harmful ideologies, manipulate political processes and reinforce social tensions. The dynamic and fast-paced nature of online media sharing behavior calls for scalable approaches to detecting malicious or misleading content. As found by Vosoughi et al. [2018], misinformation can spread considerably faster than real news. Misinformation that is repeatedly circulated has also been shown to be more persuasive to readers (Fazio et al. [2015]). While the spread of misinformation can pose risks to society regardless of intent (e.g. in the case of recent false claims about Covid-19 originating in a Chinese lab[1]), unchecked propagation of false content can pose additional harms when it is deliberately engineered to push agendas like white supremacy. In this work we address challenges in detection and mitigation of negative language use cases, e.g. misinformation and hate speech. We consider two important aspects for interpretation of perceived intent: *factuality* and *harm*. At first glance we may think the following statement generated by a large pre-trained neural language model (GPT-3, Brown et al. [2020]) is innocuous,

"I'd like to thank you.. I don't think I'd have made it if I didn't find MGTOW! You guys are great!"

[1] https://www.nature.com/articles/d41586-021-01529-3

A self-contained interpretation could be "praise for an adopted social group." However there is also pragmatic interpretations for those with knowledge of the statement's social context: MGTOW is a known misogynistic online community.

1.1 Challenges

Harmful content detectors lack awareness of communicative intent. As noted by Talat et al. [2021], machine learning models like those underlying currently deployed content moderation systems[2] are disembodied. They are trained on static data, which is separated from critical conversational and other social context (e.g. speaker vs. listener dynamics). It is an open question how much of language's true meaning can be preserved in these settings. Without social grounding which may inform perception of intent, classifiers can discriminate by ignoring nuances of dialect (Sap et al. [2019b]) or fail to identify implicit cases like microaggressions (Breitfeller et al. [2019]). An important theme of the works highlighted in this thesis is the gap between machine and human cognition with predictive systems. While state-of-the-art neural classifiers are trained to consider lexical units in context, they are prone to overweighting key terms (e.g. "terrorist"):

I want to be able to go to a country and not be looked at as a terrorist or as someone who is going to blow something up.

The statement above was given a 60% probability of toxicity by the Google Perspective API classifier. With these examples, we must consider whether neural classifiers are failing on a *semantic* level as well as a *pragmatic* level. It may be that such systems fail to understand not only language usage, but the explicit meaning of the language itself. As argued by Bender and Koller [2020], there is little reason to assume that systems trained to understand form (e.g with a next-token prediction objective) are capable of truly understanding semantics, and such failures may be just as likely to arise from incorrect or fractured explicit meaning assignment as the language model's inability to infer toxicity (or lack thereof). We use generative approaches to address these failures regardless of origin. In Chapter 3, we combine crowdsourced supervision with neural models to generate inferences about broader social context and intent of statements.

[2] `https://perspectiveapi.com/`

In Chapter 7, we propose adversarial generation approaches to address gaps in training data and prevent inequitable decision-making.

Generation evaluation is disconnected from goals. Neural models struggle with recognition of factuality as well as intent (Gabriel et al. [2021a]; Jiang and de Marneffe [2021]; Lin et al. [2021]). Generative model hallucinations arise from parameterization of models learned during training being reliant only on training data distributions, rather than real-world plausibility or verified factual knowledge. This is problematic when machine-generated text is used for tasks like fact-checking (Atanasova et al. [2020]). There have been many efforts to improve quality of generated text by aligning training objectives with factuality (Scialom et al. [2019]; Lewis et al. [2020]). However, progress in neural text generation requires having effective evaluation metrics. Existing metrics like ROUGE (Lin [2004]) and BLEU (Papineni et al. [2002]) have been shown to be incapable of measuring factuality (Gabriel et al. [2021a]; Fabbri et al. [2021]), ignore many aspects of context, and cannot identify cases of synonymous meaning. In Chapter 4, we describe a meta-evaluation that looks into whether newly proposed metrics for evaluation (e.g. question-answering based metrics) are better suited for the task of assessing factuality. We also conduct one of the first evaluations of factuality in summarization that considers a broad range of data domains (newswire and dialogue).

1.2 Methodology

Endowing neural models with pragmatic reasoning. We build systems for generating inferences about benign and harmful effects of language. The first step towards developing these systems is the conceptualization of theoretical structured formalisms for reasoning about impact and intent. These formalisms distill unstated commonsense knowledge (Gordon and Durme [2013]) relating to theory of mind (Apperly [2010]), inferring cognitive, emotional and physical effects of language on readers. Next, we translate these formalisms into supervision data for neural models using online crowdsourcing. The formalisms provide a natural structure for crowdsourcing tasks on platforms like Amazon Mechanical Turk[3], allowing real users to provide a ground-truth for questions such as "*what action might a text fragment compel a reader to take?*" Such formalisms can bridge the gap between the incomplete, unsupervised learning of social behav-

[3] https://www.mturk.com/

iors during language model pretraining through pattern recognition, and more deliberate comprehension of complexities in social dynamics. In the final stage, we train generative models (e.g. pre-trained transformer models) on human-annotated data derived using our formalism. We can transform these neural models into representations of the knowledge graph defined by our formalism through controlled generation with newly introduced "control tokens," symbols that provide a signal to the model of the dimension we are aiming to predict (e.g. perceived intent vs. invoked action). Formally, we are performing autoregressive generation (Bengio et al. [2003]) where we consider the probability of generating a token y_t, given the previously generated tokens $y_{<t}$, a context x, and a knowledge graph dimension d:

$$p(y) = \prod_{t=1}^{T} p(y_t | y_{<t}, x, d).$$

In Chapter 3, we present Misinfo Reaction Frames, a new conceptual framework and crowdsourced dataset of 69.8k unique commonsense implications for reasoning over news headlines. We show that current neural architectures can achieve high performance at detecting misinformation (up to 85.26 F1), but struggle to predict virality of news and adapt to shifting data distributions.

A critical re-evaluation of generation metrics. We uncover limitations of automated metrics and underlying models when applied to factuality evaluation in Chapter 4, by introducing a meta-evaluation for testing robustness of factuality metrics. We propose five necessary and intuitive conditions to evaluate factuality metrics on diagnostic factuality data across three different summarization tasks. The meta-evaluation involves (1) a controlled synthetic evaluation, which allows for testing the sensitivity of metrics to varying levels of factuality, and (2) an evaluation of actual machine-generated summaries to test metric robustness on the real-world distribution of errors.

Adversarial learning for improving classifier robustness. It can be challenging to surface model vulnerabilities without extensive in-the-wild testing (Kiela et al. [2021]). However, safe deployment of content moderation systems necessitates model behavior scoping and awareness of potential fairness violations. One well-studied technique for identifying model weaknesses is *adversarial example generation*, where we attack a classifier by purposefully generating text which may fool it into assigning the wrong label (Goodfel-

low et al. [2015]; Zhao et al. [2018]; Alzantot et al. [2018]). While adversarial examples have been shown to break systems for tasks like natural language inference (Glockner et al. [2018]), their efficacy for improving robustness of hate speech classifiers is considerably less studied. In Gabriel et al. [2022] and Hartvigsen et al. [2022], we show how adversarial learning can be used to mitigate discriminatory hate speech classifier behavior arising from lexical biases.

We address methods for uncovering unknown biases in Gabriel et al. [2022]. We teach generative models to mimic the behavior of classifiers by sampling lexical units that may be causally linked with the predicted label (Ribeiro et al. [2016]; Sundararajan et al. [2017]). We train the generative models to output sequences conditioned on the samples and associated label. We then substitute the opposite label during inference to encourage adversarial generation.

To prevent classifiers from associating group mentions with labels, we collect training augmentation data with a balanced distribution of group mentions for both benign and toxic examples. We use demonstration learning (Gao et al. [2021a]) with these prompt examples to generate novel text from a large language model (e.g. GPT-3, Brown et al. [2020]). We then introduce an adversarial variant of beam search decoding (Koehn [2010]) that uses a classifier-in-the-loop to control toxicity of machine-generated language. The probability from the hate speech classifier steers generations closer to the classifier's perception of toxic or benign language. By combining this with an input prompt with the opposite label of the class probability being optimized, we can push generated text closer to the classifier's decision boundary. Using this approach, we construct a large-scale dataset (274k statements) for protecting against both human- and machine-generated toxic language.

1.3 Overview of Contributions

The main contributions of this thesis are outlined as follows:

- In Chapter 2, we consider the first focus of this thesis: factuality of language. This chapter provides an introduction to the core challenges for inducing factual language production and evaluating factuality of human or machine-generated language. We also delve into what it means to tackle misinformation detection through the lens of "universal truth," and how social factors like readers' prior belief systems

complicate mitigation efforts.

- In Chapter 3, we address the challenge of representing knowledge conveyed implicitly by news claims. We describe a pragmatic formalism for inferring reader reactions to news headlines. We describe how this formalism can be used to reason about the effects of claims on social behavior, including cognitive and physical effects. We also show how data collected using this formalism can be used as supervision for a neuro-symbolic approach to reasoning over news claims. This allows us to take initial steps towards constructing probabilistic models of human behavior for better analyzing effects of misinformation. This work was published previously in Gabriel et al. [2021b].

- In Chapters 4 and 5, we address foundational issues in evaluation of text generation which present a bottleneck to use of machine-generated counternarratives. We introduce a theoretically grounded meta-evaluation to measure the effectiveness of commonly used generation metrics at assessing factuality. We confirm that current metrics are often insufficient for evaluating factuality, and began to explore the dangers of over-generalizing metric performance with a medical data case study. The majority of this work was published previously in Gabriel et al. [2021a].

- In Chapter 6, we transition to the broader topic of harms posed by both human- and machine-generated text. Large language models have led to a new threat landscape that exacerbates many existing societal risks like identity theft and dissemination of hate speech. However, such models are exemplars of dual-use technologies in that they can also function as defenses against harmful content (Zellers et al. [2019]). We provide background for how prior work has handled ethical quandaries raised by controlled generation for purposes of harm mitigation.

- In Chapter 7, we show that large pretrained neural language models like GPT-3 can operate like retrieval models for dangerous web data. Not only can they be easily induced through prompting to generate non-factual content, they also generate racist, sexist and otherwise inflammatory content that concerningly mirrors human-written hate speech. However, this behavior can be controlled by the chosen input structure and decoding method. We introduce two approaches that provide us with the capability of generating novel, hard examples that fool existing classifiers. We also show that this form of data generation can be used to retrain classifiers to be considerably more robust. This work

was previously published in Hartvigsen et al. [2022] and Gabriel et al. [2022].

- In Chapter 8, we discuss initial steps towards developing an universal text-and-vision model for detection of potential harms. This work consolidates findings from our earlier studies into a framework for social grounding and continual improvement of classifiers.

Through these works, we argue for pragmatic approaches to detection of misinformation and toxicity. Severing the connection between AI systems and the nuances of social contexts in which they are used, effectively disemboding such models, risks discriminating against the same identity groups commonly targeted by harmful content like hate speech (Sap et al. [2019b]). In contrast, pragmatic approaches like those we propose can adapt to shifting social contexts. We conclude with a vision for community-driven NLP, which seeks to not only mitigate in a posthoc manner the harms inflected by AI, but instead proactively prevent such harms through better-informed algorithm design. We explore the data-centric, algorithmic and structural reforms needed for this research philosophy, as well as how this may address known challenges within the field like distributional robustness.

Part I

Factuality

Chapter 2

The Search for Universal Truth

Misinformation mitigation efforts have focused on a knowledge deficit model (Simis et al. [2016]; Ecker et al. [2022]), in which it is assumed misinformation can be overcome by improving access to facts. This model motivates countermeasures like social media warning labels on posts flagged by fact-checkers[1][2] and other forms of content indicators (Zhang et al. [2018a]). However this model relies on the assumption of users as being rational agents, who can agree upon a common "universal" truth once exposed to enough information. Factual information can have a limited effect on users if their prior beliefs reinforce their trust in false content (Chambliss and Garner [1996]; Hart et al. [2009]). In a striking example of ideological entrenchment, Bail et al. [2018] found that exposing users to information from an opposing political ideology can worsen polarization.

People who hold strong opinions on complex social issues are likely to examine relevant empirical evidence in a biased manner. They are apt to accept "confirming" evidence at face value while subjecting "disconfirming" evidence to critical evaluation, and as a result to draw undue support for their initial position from mixed or random empirical findings. - From Lord et al. [1979]

Given that users are prone to confirmation bias (Nickerson [1998]), we need to understand the mental models of those users in order to counteract misinformation. This requires methods for representing and communicating the state of the world according to a potential user, U. One strategy is to take a Bayesian

[1] https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news
[2] https://help.twitter.com/en/resources/addressing-misleading-info

approach (Frank and Goodman [2012]; Acemoglu et al. [2022]), defining a prior world state θ from a user-centric perspective based on the ideology of U. In the next chapter we describe how we model the cognitive processes of users using neural networks to predict a probability distribution over potential reactions to news claims. While our conceptualization normalizes over users, it can be adapted to condition on the beliefs of a particular user U_i and provide ideologically informed inference.

2.1 What is misinformation?

Both disinformation and misinformation are defined as misleading or false content, however disinformation assumes malicious authorial intent.[3] In this thesis, we focus on the reader's perception of the writer's intent. This is due to the challenges of recovering original authorial intent in collapsed contexts prevalent on social media (Starbird et al. [2019]). We summarize common definitions for news reliability below:

- ***Misinformation*** is an umbrella term for news that is false or misleading. It assumes accidental rather than malicious propagation.

- Unlike misinformation, ***disinformation*** assumes a malicious intent or desire to manipulate (Fallis [2014]).

- As defined by Allcott and Gentzkow [2017], ***fake news*** refers to *"news articles that are intentionally and verifiably false, and could mislead readers."* Golbeck et al. [2018] notes that fake news is a form of hoax, where the content is factually incorrect and the purpose is to mislead. This also overlaps with the definition of disinformation.

- ***Propaganda*** is widely held to be news that is *"an expression of opinion or action by individuals or groups, deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends"* (Miller [1939]). Propaganda is therefore wholly defined in terms of the intent of a writer or group of writers, and may contain factually correct content.

- We refer to articles written with a humorous or ironic intent as ***satire***. We do not explicitly cover satire in this thesis, but it is possible that misinformation articles can begin as satire and be misconstrued as

[3] `https://www.un.org/en/countering-disinformation`

real news.

- ***Verified news*** is considered to be news that has been confirmed by fact-checkers to be factually correct with an intent to inform. We note that while real news is distinct from most of the article types shown here, it can also function as propaganda. The findings of Gabriel et al. [2021b] show that a substantial portion of modern news articles are predicted to contain propagandistic language by neural rhetorical technique detectors.

2.2 How is misinformation detected?

The general setting of the misinformation detection or fact-checking task is to define a function that given a claim C can map it to a reliability label $\hat{y}$ (e.g. C $\rightarrow \hat{y} \in \{\text{misinfo}, \text{real}\}$). There can be variation in the granularity of the label space (Augenstein et al. [2019]), and fact-checking approaches use retrieved textual and/or visual evidence E for verification rather than solely relying on parametric knowledge (Vlachos and Riedel [2014]).

Determination of Factuality We generally leave the process of verifying gold labels to external professional fact-checkers. Labels are typically determined on either an article or source level. In the former case, articles are individually examined for misleading content by journalistic organizations like Poynter.[4] This is the most effective approach since it can uncover mixed reliability articles or errors by trusted news organization, but available datasets are limited because of the time commitment. In the latter case, articles can automatically be assigned a label based on a database of trusted and suspicious news media (Nørregaard et al. [2019a]). It has been found that layperson crowdsource workers can be effective fact-checkers under the right conditions (Allen et al. [2021]), motivating community-driven approaches to mitigating misinformation.

Automated Detection Given the rapid proliferation of misinformation on social media, there is a critical need for automatic tools that can assist human fact-checkers (Nakov et al. [2021]). Much of the earlier work in this area relied on linguistic cues or social media network features (e.g, Wang [2017]; Rashkin et al.

[4] https://www.poynter.org/ifcn/

[2017]; Pérez-Rosas et al. [2018]; Yang et al. [2019]). Later work on detection of mis- and disinformation has considered transformer-based approaches, notably (Zellers et al. [2019]; Lee et al. [2021]). Prior to us, Angeli and Manning [2014] draws a connection between claim verification and the natural logical inference underlying commonsense acquisition.

Chapter 3

Pragmatic Frames of News

In order to identify inflammatory content, prevent discriminatory behavior of AI systems and ground such systems with nuanced social context, we must first be able to construct a predictive model of how textual information may be perceived by readers. In this chapter, we introduce an approach for actualizing such predictive models through crowdsourced knowledge acquisition and neuro-symbolic reasoning. We focus on a case study of predicting how Covid-19 or Climate news headlines may influence readers (Gabriel et al. [2021b]).

While most prior NLP research on misinformation has focused on fact-checking, preventing spread of misinformation goes beyond determining veracity (Schuster et al. [2019]; Ren et al. [2023]). For example, mistrust in the government may lead readers to share pandemic conspiracy headlines like *"Epidemics and cases of disease in the 21st century are "staged""* even if they suspect it is misinformation. This task requires knowledge of how diverse readers perceive the intent behind real and fake news. This is extremely challenging given that headlines often make use of implicit messaging that recalls broader social context or world knowledge, making prior information extraction approaches (Andersen et al. [1992]; Agichtein and Gravano [2000]; Etzioni et al. [2008]) less effective for reasoning about intent. We introduce **Misinfo Reaction Frames** (MRF), a pragmatic formalism to reason about the effect of news headlines on readers. Inspired by Frame Semantics (Fillmore [1976]), our frames distill the pragmatic implications of a news headline in a structured manner. We capture free-text explanations of readers' reactions and perceived author intent, as well as categorical estimates of veracity and likelihood of spread.

3.1 Misinfo Reaction Frames

3.1.1 Motivation: Challenges in Determining Intent

In contrast to prior work on misinformation detection (Ott et al. [2011]; Rubin et al. [2016]; Rashkin et al. [2017]; Wang [2017]; Hou et al. [2019]; Volkova et al. [2017]; Jiang and Wilson [2018]) which mostly focuses on linguistic or social media-derived features, we focus on the potential impact of a news headline by modeling readers' reactions. This can help us counter misinformation given the need for a mental model of users (see Chapter 2 for the full motivation). It has also been shown that interventions from AI agents are better at influencing readers than strangers (Kulkarni and Chi [2013]).

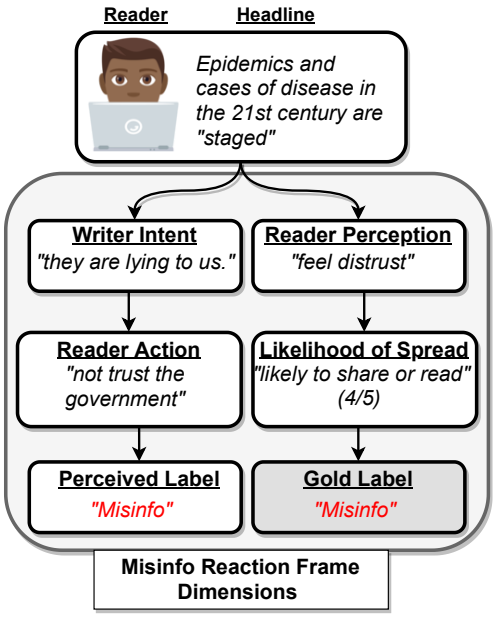

Figure 3.1: Our pragmatic frames (Misinfo Reaction Frames) explain how a news headline is interpreted as reliable or misinformation by readers.

In order to model impact, we build upon prior work that aims to describe the rich interactions involved in human communication, including semantic frames (Fillmore [1976]), the encoder-decoder theory of media (Hall [1973]),[1] Grice's conversational maxims (Grice [1975]) and the rational speech act model (Goodman and Frank [2016]).[2] By describing these interactions with free-text implications invoked by a news headline, we also follow from prior work on pragmatic frames of connotation and social biases (Speer and Havasi [2012]; Rashkin et al. [2018]; Sap et al. [2019a, 2020]; Forbes et al. [2020]). We note that foundational prior work on pragmatic understanding of language assume cooperative speaker behavior (Goodman and Frank [2016]), which may not fit a world model in which the speaker is purposefully malicious, obscure or misleading. By bridging communication theory, data annotation schema and predictive modeling, we define a concrete framework for understanding the impact of either real news or misinformation headlines on a reader.

[1] This theory proposes that before an event is communicated, a narrative discourse encoding the objectives of the writer is generated.

[2] Here pragmatic interpretation is framed as a probabilistic reasoning problem.

Theme	Climate	Covid
Climate Statistics	✓	
Natural Disasters	✓	
Entertainment	✓	
Ideology	✓	
Disease Transmission		✓
Disease Statistics		✓
Health Treatments		✓
Protective Gear		✓
Government Entities	✓	✓
Society	✓	✓
Technology	✓	✓

Table 3.1: Themes present in articles by each news topic. Some are covered by both climate and Covid domains, while others are domain specific.

Statistic	Train	Dev.	Test
Headlines	19,897	2,460	2,133
Intents	38,172	4,867	4,388
Percept.	2,609	538	421
Actions	15,036	2,176	1,739
Total Pairs	159,564	19,700	17,890

Table 3.2: Dataset-level breakdown of headlines, as well as unique and total implications for MRF corpus.

3.1.2 Defining a Taxonomy for Intent and Impact

In this section, we define the structure of the Misinfo Reaction Frames formalism and detail how we use the structured formalism to construct a corpus of commonsense implications aligned with news headlines (§3.2). Each reaction frame contains the dimensions in Figure 3.1. We elicit annotations based on a *news headline*, which summarizes the main message of an article. An example headline is "*Covid-19 may strike more cats than believed*." To simplify the task for annotators and ground implications in real-world concerns, we define these implications as relating to one of 7 common themes (e.g. technology or government entities) appearing in Covid and climate news.[3] We list all the themes in Table 3.1, with some themes being shared between topics.

3.2 The Reaction Frames Corpus

To construct a corpus for studying reader reactions to news headlines, we obtain 69,885 news implications by eliciting annotations for 25,164 news headlines (11,757 Covid related articles, 12,733 climate headlines and 674 cancer headlines). There are two stages for collecting the corpus - (1) news data collection described in §3.2.1 and (2) crowd-sourced annotation described in §3.2.2.

[3]We use a subset of the data (approx. 200 examples per news topic) to manually identify themes. Note that themes are not disjoint and a news article may capture aspects of multiple themes.

3.2.1 News Data Collection

As described in Chapter 2, a number of definitions have been proposed for labeling news articles based on reliability. To scope our task, we focus on false news that may be unintentionally spread (misinformation). We examine reliable and unreliable headline extracted from two domains with widespread misinformation: Covid-19 (Hossain et al. [2020]) and climate change (Lett [2017]). We additionally test on cancer news (Cui et al. [2020]) to measure out-of-domain performance.

Climate Change Dataset We retrieve both trustworthy and misinformation headlines related to climate change from NELA-GT-2018-2020 (Nørregaard et al. [2019a,b]), a dataset of news articles from 519 sources. Each source in this dataset is labeled with a 3-way trustworthy score (reliable / sometimes reliable / unreliable). We discard articles from "sometimes reliable" sources since the most appropriate label under a binary labeling scheme is unclear. To identify headlines related to climate change, we use keyword filtering.[4] We also use claims from the SciDCC dataset (Mondal et al. [2021]), which consists of 11k real news articles from ScienceDaily,[5] and Climate-FEVER (Diggelmann et al. [2020]), which consists of more than 1,500 true and false climate claims from Wikipedia. We extract claims with either supported or refuted labels in the original dataset.[6]

Covid-19 Dataset For trustworthy news regarding Covid-19, we use the CoAID dataset (Cui and Lee [2020]) and a Covid-19 related subset of NELA-GT-2020 (Nørregaard et al. [2019a]). CoAID contains 3,565 news headlines from reliable sources. These headlines contain Covid-19 specific keywords and are scraped from nine trustworthy outlets (e.g. the World Health Organization).

For unreliable news (misinformation), we use The CoronaVirusFacts/DatosCoronaVirus Alliance Database, a dataset of over 10,000 mostly false claims related to Covid-19 and the ESOC Covid-19 Misinformation Dataset, which consists of over 200 additional URLs for (mis/dis)information examples.[7][8] These claims originate from social media posts, manipulated media, and news articles, that have been manually reviewed

[4] We kept any article headline that contained at least one of "environment," "climate," "greenhouse gas," or "carbon tax." We remove noisy examples obtained using these keywords with manual cleaning.

[5] `https://www.sciencedaily.com/`

[6] The data also includes some claims for which there is not enough info to infer a label. We discard these claims.

[7] `https://www.poynter.org`

[8] `esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset`

and summarized by fact-checkers.

Cancer Dataset We construct an evaluation set for testing out-of-domain performance using cancer real and misinformation headlines from the DETERRENT dataset (Cui et al. [2020]), consisting of 4.6k real news and 1.4k fake news articles.

3.2.2 Annotation Process

In this section we outline the structured annotation interface used to collect the dataset. Statistics for the dataset are provided in Table 3.2.

Annotation Task Interface We use the Amazon Mechanical Turk (MTurk) crowdsourcing platform.[9] For ease of readability during annotation, we present a headline summarizing the article to annotators, rather than the full text of the article. Annotators then rate veracity and likelihood of spread[10] based on the headline, as well as providing free-text responses for writer intent, reader perception and reader action.[11]

Quality Control We use a three-stage annotation process for ensuring quality control. Qualified workers are located in the US, have had at least 99% of their *human intelligence tasks* (hits) approved and have had at least 5000 hits approved. We removed workers whose accuracy at predicting the label (real/misinfo) of news headlines fell below 70%. We achieve pairwise agreement of 79% on the label predicted by annotators during stage 3. To account for chance agreement, we also measure Cohen's Kappa $\kappa = .51$, which is considered "moderate" agreement.

3.3 Modeling Reaction Frames

We test the ability of large-scale language models to predict Misinfo Reaction Frames. For free-text inferences (e.g. writer intent, reader perception), we use generative language models, specifically T5 encoder-decoder (Raffel et al. [2020]) and GPT-2 decoder-only models (Radford et al. [2019]). For categorical

[9] https://www.mturk.com/

[10] This is based on their perception of likely virality and personal preferences.

[11] These news events are either article headlines or claims.

inferences (e.g. the gold label), we use either generative models or BERT-based discriminative models (Devlin et al. [2019]). We compare neural models to a simple retrieval baseline (**BERT-NN**) where we use gold implications aligned with the most similar headline from the training set.[12] We evaluate reaction inference systems using common automatic metrics (§3.3.4). We also use human evaluation to assess quality and potential use of generated writer intent inferences (§7.2.6).

3.3.1 Controlled Generation

For generative models, we use the following input sequence

$$x = h_1 \ ... \ h_T || s_d || s_t,$$

where h is a headline of length T tokens, $s_t \in \{$[*covid*],[*climate*]$\}$ is a special topic control token, and s_d is a special dimension control token representing one of six reaction frame dimensions. Here $||$ represents concatenation. The output is a short sequence representing the predicted inference (e.g. "*to protest*" for reader action, "*misinfo*" for the gold label). For GPT-2 models we also append the gold output inference $y = g_1 \ ... \ g_N$ during training, where N is the length of the inference.

Inference We predict each token of the output inference starting from the topic token s_t until the $[eos]$ special token is generated. In the case of data with unknown topic labels, this allows us to jointly predict the topic label and output inference. We decode using beam search, since generations by beam search are known to be less diverse but more factually aligned with the context (Massarelli et al. [2020a]).

3.3.2 Classification

For discriminative models, we use the following input sequence

$$x = [CLS] h_1 \ ... \ h_T [SEP],$$

where [CLS] and [SEP] are model-specific special tokens. The output is a categorical inference.

[12] Similarity is measured between headlines embedded with MiniLM, a distilled transformer model (Wang et al. [2020b]). We use the Sentence-BERT package (Reimers and Gurevych [2019]).

3.3.3 Training

All our models are optimized using cross-entropy loss, where generally for a sequence of tokens t

$$CE(t) = -\frac{1}{|t|}\sum_{i=1}^{|t|} logP_\theta(t_i|t_1, ..., t_{i-1}).$$

Here P_θ is the probability given a particular language model θ.

3.3.4 Automatic Metrics

These metrics include the BLEU (-4) ngram overlap metric (Papineni et al. [2002]) and BERTScore (Zhang et al. [2020]), a model-based metric for measuring semantic similarity between generated inferences and references. For classification we report macro-averaged precision, recall and F1 scores.[13][14] We use publicly available implementations for all metrics (nltk[15] for BLEU). Automatic results are omitted here for brevity, and can be found in the original paper.

3.3.5 Human Evaluation

For human evaluation, we assess generated inferences using the same pool of qualified workers who annotated the original data. We randomly sample model-generated "writer's intent" implications from T5 models and GPT-2 large over 196 headlines where generated implications were unique for each model type.[16] We elicit 3 unique judgements per headline. Implications are templated in the form *"The writer is implying that [implication]"* for ease of readability.

Overall Quality We ask the annotators to assess the overall quality of generated implications on a 1-5 Likert scale (i.e. whether they are coherent and relevant to the headline without directly copying).

Influence on Trust We measure whether generated implications impact readers' perception of news reliability by asking annotators whether a generated implication makes them perceive the news headline as more

[13] We compute these using scikit-learn: `https://scikit-learn.org/stable/index.html`
[14] For measuring likelihood of spread, predicted and averaged values are rounded to the nearest integer.
[15] `https://www.nltk.org/`
[16] 98 misinfo and 98 real headlines in the dev. set

(+) or less (-) trustworthy.

Perceived Sociopolitical Acceptability We ask annotators to rate their perception of the beliefs invoked by an implication in terms of whether they represent a majority (mainstream) or minority (fringe) viewpoint.[17]

A/B Testing For A/B testing, annotators are initially shown the headline with the generated implication hidden. We ask annotators to rate trustworthiness of headlines on a 1-5 Likert scale, with 1 being clearly misinformation and 5 being clearly real news. After providing this rating, we reveal the generated implication to annotators and have them rate the headline again on the same scale. Annotators were not told whether or not implications were machine-generated, and we advised annotators to mark generated implications that were copies of the headlines as low quality.

3.3.6 Results

		Writer Intent		Reader Perception		Reader Action	
	Model	BLEU-4 ↑	BERTScore ↑	BLEU-4 ↑	BERTScore ↑	BLEU-4 ↑	BERTScore ↑
dev.	BERT-NN	31.45	86.29	35.69	91.04	45.47	84.76
	T5-base	51.48	88.03	31.98	92.87	53.55	85.27
	T5-large	51.30	**88.16**	32.82	**92.94**	57.29	**85.34**
	GPT-2 (small)	**60.68**	87.35	**37.22**	92.21	54.20	84.83
	GPT-2 (large)	54.94	87.74	32.35	92.84	**57.84**	85.00
test	BERT-NN	34.46	86.35	**37.09**	90.84	46.57	84.78
	T5-base	50.63	87.78	32.18	**93.32**	57.37	85.60
	T5-large	50.86	**87.94**	32.89	93.29	**62.10**	**85.88**
	GPT-2 (large)	**60.51**	87.73	34.18	92.51	59.57	85.53

Table 3.3: Automatic modeling results (generation task). For this table and the following tables, we highlight the best-performing model(s) in **bold**.

Generating Reaction Frames

The automatic evaluation results of our generation task are provided in Table 3.3. Human evaluation results of our generation task are provided in Table 3.4.

[17] We refer to “minority" viewpoint broadly in terms of less frequently adopted or extreme social beliefs, rather than in terms of viewpoints held by historically marginalized groups.

Model	Quality (1-5)	Influence on Trust +Trust (%)	-Trust (%)	Corr w/ Label (all gens)	Corr w/ Label (quality ≥ 3)	Socially Acceptable? (%)
T5-base	3.61	8.33	7.82	**0.24***	**0.30***	**75.30**
T5-large	**3.74**	7.73	9.76	-0.03	0.09	74.66
GPT-2 (large)	3.46	**9.70**	**13.10**	-0.04	0.10	74.66

Table 3.4: Human evaluation results (generation task). Cells marked by "*" are statistically significant for $p < .05$.

Results We found that the T5-large model was rated as having slightly higher quality generations than the other model variants (Table 3.4). Most model generations were rated as being *"socially acceptable"*. However in as many as 25.34% of judgements, generations were found to be not socially acceptable.

Interestingly, all models were rated capable of influencing readers to trust or distrust headlines, but effectiveness is dependent on the quality of the generated implication. In particular for T5-base, we found a consistent correlation between the actual label and shifts in trustworthiness scores before and after annotators see the generated writer's intent. Annotators reported that writer intents made real news appear more trustworthy and misinformation less trustworthy.[18]

Detecting Misinformation

To test if we can detect misinformation using propagandistic content like *loaded or provocative language* (e.g. "Covid-19 vaccines may be *the worst threat we face*"), we use a pre-trained BERT propaganda detector (Martino et al. [2019]) which we denote here as (Prop-BERT).[19] For our zero-shot setting, we classify a news event as real if it is not associated with any propaganda techniques and misinformation otherwise. As shown by Table 3.5, F1 results are considerably lower than task-specific models. This is likely due to the fact both real and misinformation news uses propaganda techniques.

Neural misinformation detection models are able to outperform humans at identifying misinformation (achieving a max F1 of 85.26 compared to human performance F1 of 75.21[20]), but this is still a nontrivial task for large-scale models. When we use Covid-BERT (Müller et al. [2020]), a variant of BERT pretrained

[18]While for most models the trend is a decrease in trust for both real news and misinformation, for the T5-base model there is a statistically significant correlation of Pearson's $r = .24$ showing shifts in trust align with gold labels.

[19]The model predicts if any of 18 known propaganda techniques are used to describe a news event. See the paper for the full list.

[20]We count disagreements as being labeled misinformation here, discarding disagreements leads to F1 of 74.97.

	Model	Spread F1 ↑	Gold F1 ↑
dev.	Majority Baseline	10.49	34.49
	T5-base	22.77	87.13
	T5-large	**29.04**	88.12
	GPT-2 (small)	22.38	83.86
	GPT-2 (large)	27.59	89.01
	Prop-BERT	-	46.43
	BERT-large	-	89.24
	Covid-BERT	-	**90.60**
test	Majority Baseline	11.20	38.58
	T5-base	20.59	80.43
	T5-large	**30.60**	81.20
	GPT-2 (large)	18.41	81.35
	Prop-BERT	-	38.79
	BERT-large	-	79.80
	Covid-BERT	-	**85.26**
cancer (unsup.)	Prop-BERT	-	61.78
	BERT-large	-	30.07
	Covid-BERT	-	56.85
	GPT-2 (large)	10.95	43.50
	T5-large	21.12	35.52
	GPT-2 (large) + masked	**21.78**	**65.99**
	T5-large + masked	19.57	45.91

Table 3.5: Automatic modeling results (classification task). The spread variable models the likelihood of news being read or shared, while gold indicates the fact-checked label (real/misinfo). For the unsupervised cancer setting (unsup.), all models are trained on covid/climate data only or another news dataset (Prop-BERT). For the unsupervised setting (unsup.), we evaluate on 100 cancer news examples.

on 160M Covid-related tweets, we see an improvement of 5.46% over BERT without domain-specific pre-training (Table 3.5). This indicates greater access to domain-specific data helps in misinformation detection, even if the veracity of claims stated in the data is unknown.

Performance on Out-of-Domain Data

We test the ability of reaction frames to generalize using 100 cancer-related real and misinformation health news headlines (Cui et al. [2020]), see Table 3.5. To improve generalization of MRF models, we use an additional masked fine-tuning step. We first train a language model θ on a set of Covid-19 training examples D_{covid} and climate training examples $D_{climate}$. Then we use the Textrank algorithm (Mihalcea and Tarau [2004]) to find salient keyphrases in D_{covid} and $D_{climate}$, which we term k_{covid} and $k_{climate}$ respectively.

We determine domain-specific keyphrases by looking at the complement of $k_{covid} \cap k_{climate}$ and only keep the top 100 keyphrases for each domain. We mask out these keyphrases in the training examples from D_{covid} and $D_{climate}$ by replacing them with a $< mask >$ token. Then we continue training by fine-tuning on the masked examples. We denote these models in Table 3.5 as "+ masked." For the misinformation detection task, we evaluate gold F1 using the Prop-BERT zero-shot model, MRF-finetuned BERT-large, Covid-BERT, T5-large and GPT-2 large models. We observe that after one epoch of re-training, masked fine-tuning substantially boosts unsupervised performance of generative MRF models (GPT-2 large + masked and T5-large + masked), making them more robust than BERT variants.

3.3.7 Summary of Key Findings

Our framework presents new opportunities for studying perceived intent and impact of misinformation, which may also aid in countering and detecting misinformation:

We can estimate content virality. Given the user-annotated labels for likelihood of reading or sharing, we can estimate whether the information in the associated article is likely to propagate.

We can analyze the underlying intents behind headlines. Using annotated writer intents, we can determine common themes and perceived intentions in misinformation headlines across domains (e.g. mistrust of vaccination across medical domains). Given the performance of predictive models highlighted by Tables 3.3 and 3.4, we can also extend this analysis to unseen headlines.

We can categorize headlines by severity of likely outcomes. False headlines that explicitly incite violence, or otherwise encourage actions that lead to psychological or physical harm (e.g. not vaccinating) may be deemed more malicious than false headlines with more benign consequences (e.g. some examples of satire). Future work may explore categorizing severity of headlines based on potential harms resulting from implications.

Perceived labels can help us understand which headlines may fool readers. We can use these labels to determine which types of misinformation headlines appear most like real news to generally knowledge-

able readers. These may also help in designing misinformation countering systems and better adversarial examples to improve robustness of misinformation detection models.

We can generate counter-narratives to misinformation. Our results indicate it is possible to generate effective explanations for the intent of headlines that discourage trust in misinformation (Section 3.3.6).

Limitations. Given these future directions, we also consider key limitations which must be addressed if we move beyond viewing Misinfo Reaction Frames as a proof-of-concept and use the dataset as part of a large-scale system for evaluating or countering misinformation.

- Since we focus on news headlines, the context is limited. The intent of a headline may be different from the actual intent of the corresponding article, especially in the case of clickbait. We find headlines to be suitable as online readers often share headlines without clicking on them (Gabielkov et al. [2016]), however future work may explore extending reaction frames to full news articles.

- There is also annotator and model bias. Readers involved in our data curation and human evaluation studies are "*generally knowledgeable*," as proved by their ability to discern misinformation from real news. We see this bias as a potential strength as it allows us to find ways to counter misinformation in cases where readers are well-informed but still believe false information. However, annotators may have undesirable political or social biases. In such cases, gender bias may lead an annotator to assume that a politician mentioned in a headline is male or to dismiss inequality concerns raised by a scientist belonging to a minority group as "playing the race card." These biases can also appear in pre-training data, leading to model bias.[21] Subjectivity in annotation is a point of discussion in many pragmatic-oriented tasks, e.g. social norm prediction Jiang et al. [2022] and toxicity detection Halevy et al. [2021]; Sap et al. [2022a]. We encourage conscious efforts to recruit diverse pools of annotators so multiple perspectives are considered, and future work on modeling reaction frames can consider learning algorithms that mitigate harmful effects of biases, depending on use case Khalifa et al. [2021]; Gordon et al. [2022].

[21] Removing these examples from data curation or trying to control for "annotator neutrality" does not erase the causes that lead to the existence of these biases. The fact that harmful biases can manifest in the viewpoints of informed readers speaks to the pervasiveness of certain stereotypes.

- We only consider English-language news and annotate with workers based in the US. It may be that news headlines would be interpreted differently in other languages and cultures.
- Lastly, use of machine-generated explanations for news intent is challenged by the propensity of neural language models to hallucinate (Tian et al. [2019]; Dziri et al. [2021]). In the next chapter, we directly address these challenges and introduce testing strategies for finding evaluation metrics that can measure factual consistency.

Chapter 4

Factuality Meta-Evaluation

The goal of text generation systems is to produce text that is fluent, coherent, relevant, as well as factually correct. Recent progress in neural approaches to building semantically constrained text generation systems has shown tremendous improvements in this direction (Liu and Lapata [2019]; Guo et al. [2018]; Durmus et al. [2020]; Wang et al. [2020a]). However, an important issue in text generation systems is that they can yield factually inconsistent text, caused by somewhat distorted or fabricated facts about the source text. Especially in document summarization tasks, models that abstract away salient aspects have been shown to generate factual inconsistencies (Kryscinski et al. [2020]; Falke et al. [2019a]; Zhu et al. [2021]).

Commonly used metrics for measuring quality of generated text fail to capture structural aspects of language like negation (see Figure 4.1) and poorly correlate with human judgements (Hashimoto et al. [2019]; Clark et al. [2019]; Sellam et al. [2020]), leading to a rapidly progressing search for factuality-driven summarization metrics.

4.1 Related Work on Evaluation

Prior work concerning evaluation of automatic metrics and human evaluation for NLG systems has mainly focused on general analysis of output quality or coherence and fluency (Callison-Burch et al. [2007]; Graham [2015]; Fabbri et al. [2020]), rather than factuality. Recent efforts by NLP researchers have drawn attention to the issue of factual errors and hallucinations in the output of neural summarization models (Cao et al. [2018]; Massarelli et al. [2020b]; Falke et al. [2019b]; Goodrich et al. [2019]; Zhao et al. [2020]; Celikyilmaz

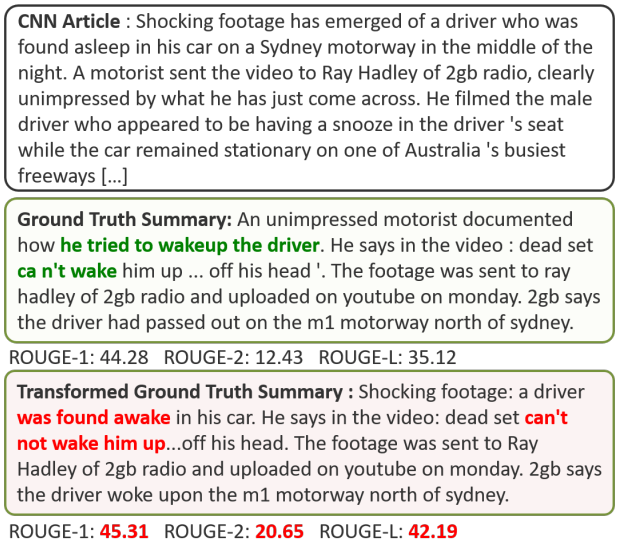

Figure 4.1: Example of a ground-truth CNN/DailyMail summary and transformed summary where key spans of the ground-truth summary (highlighted in green) contain factual errors (highlighted in red). Even though the transformed summary is less factual, the commonly used ROUGE summarization metric assigns higher values to that summary over the ground-truth summary when we compare against the original article as a reference.

et al. [2020]). A number of works have highlighted the effectiveness of QA and cloze task objectives for evaluating or improving factuality on specific domains (Eyal et al. [2019]; Huang et al. [2020]). We aim to evaluate these metrics more broadly, and consider a wider range of domains (notably dialogue).

4.2 Factuality Metrics for Evaluation

We mainly focus on meta-evaluating most recently proposed factual consistency metrics which use proxy natural language understanding (NLU) objectives aimed at implicitly capturing factuality in generated text:

QA-Based Quality Score. Given a source or reference document D and candidate summary S_i, QA-based evaluation metrics assign a generation quality score to S_i to measure the ability of a QA system by accurately answering questions generated from D or S_i. We use the SummaQA (Scialom et al. [2019]) and FEQA (Durmus et al. [2020]) metrics. For the SummaQA metric, questions are generated from the source document D and the candidate summary S_i is used as input to the QA system. Alternatively, FEQA generates questions from S_i and uses D to answer these questions.

The generation quality score is typically the aggregated F_1 score measuring the similarity between

ground-truth answers for questions generated from D and the answers predicted by the QA system. SummaQA also generally includes the aggregated model confidence probabilities for predictions.

Masked LM Prediction (Cloze Task) Score. Given a source document D and candidate summary S_i, Cloze-based evaluation metrics assign a generation quality score to S_i by measuring the ability of a NLU system to accurately predict masked tokens in the source document, given access to the information in S_i. We use two variants of BLANC (Vasilyev et al. [2020]), BLANC-Help and BLANC-Tune. BLANC-Help uses both D and S_i as input to a pretrained masked token prediction model, while BLANC-Tune only uses D as input to a model that has been finetuned on the candidate summary. Both metrics are aimed at capturing fluency, informativeness and factual correctness of summaries.

Semantic Similarity. Semantic similarity metrics measure the overlap between contextual embeddings of a source or reference document D and candidate summary S_i. We use BERTScore (Zhang et al. [2020]), which has been shown to correlate better with human judgements of coherency than standard summarization metrics and similarly to *n*-gram metrics on factual consistency of CNNDM summaries (Wang et al. [2020a]).

4.3 Defining Desired Metric Attributes

Since reference summaries may be an incomplete representation of the salient facts in a source document or unavailable, we consider factuality in terms of how well candidate summaries are factually grounded with respect to the source document.

We define a set of five conditions for a factual consistency metric $M(D, S_i)$ to measure factuality of a summary S_i with respect to a source document D. These conditions are given in Table 4.1.

4.3.1 Testing Factuality Metric Validity

For the purposes of testing boundedness (Condition I), we define the **Lower Bound** for a metric M as $M(D, S_r)$ where D is the source document and S_r is a randomly sampled summary from the corpus.[1] We

[1]While this may not be the strictest lower bound in theoretical terms, we consider it appropriate as an empirical lower bound since the content is irrelevant to the document. A single random summary is used.

Condition	Definition	Motivation
Boundedness (I)	There exists S_r, S_f such that $M(D, S_r) \leq M(D, S_i) \leq M(D, S_f)$.	In general, the exact factuality level of S_i may be unclear. Metric bounds provide points of comparison.
Sensitivity (II)	The metric value for S_i should correlate with the level of factuality captured by S_i.	A bounded but insensitive factuality metric may assign higher values to mostly nonfactual or unrelated summaries over summaries that are close to the reference.
Robustness (III)	The metric should be *robust* across types of factual errors.	A metric that is sensitive only to a subset of errors might ignore a significant number of model-generated errors.
Generality (IV)	The metric should satisfy conditions I,II,III and V across domains.	Prior work such as Reiter and Belz [2009] highlight the risk of claiming validity without testing generality.
Human Correlation (V)	The metric should *correlate* with human judgements of factuality.	The scoring function $H(D, S_i)$ represented by human evaluation is a gold standard for assessment of generation quality Chaganty et al. [2018], so $M(D, S_i)$ should be an approximation.

Table 4.1: Details of factuality metric conditions. Here M is a metric scoring function, D is a source document and S_i is a summary.

define the **Upper Bound** for the metric as $M(D, S_f)$, where S_f is the reference ground-truth summary. Since our controlled experiments use transformed versions of the reference summary with injected errors, the original reference is guaranteed to be at least as factually consistent as a transformed summary.

To test sensitivity (Condition II), we measure the **correlation** (Pearson's r) between the factual inconsistency level[2] of the summaries (i.e. the number of injected errors) and the average metric score. Then we measure statistical significance using the ***p*-value** from a two-tailed hypothesis test. We check whether metrics satisfy robustness and generality (Conditions III and IV) by separately running this analysis over multiple domains and factual error types. We measure how well metric values match human assessment of factuality by checking the correlation between factual consistency levels determined using manual annotation.

[2]For our experiments, we inject up to a maximum of x errors with $x \in \{1, 2, 3\}$.

4.4 Experimental Setup

We evaluate metrics on three datasets: 1-sentence BBC news summaries from the XSUM extreme summarization dataset (Narayan et al. [2018]), multi-sentence summaries from the CNN/DailyMail dataset (Nallapati et al. [2016]), and the recently released SAMSUM corpus (Gliwa et al. [2019]) consisting of English language conversations written by linguists and aligned multi-sentence summaries.

4.4.1 Diagnostic Datasets

To test the ability of proposed metrics to fulfill our predefined conditions, we set up two diagnostic datasets consisting of (*i*) transformed reference summaries with simulated factuality errors that allow us to induce and measure factuality levels in a controlled setting and (*ii*) summaries generated by state-of-the-art transformer summarization models that allows us to measure the effectiveness of metrics in a real data setting. We sample 500 source / summary pairs for each domain.

Controlled Error Setting

We inject errors into reference summaries by first using a part-of-speech tagging model and named entity recognition system (spaCy)[3] to extract entities, verbs, and adjectives from these summaries. For each named entity, we keep track of the label type (e.g. ORG, GPE, etc). All datasets are comprised of English language articles or dialogues and summaries, and we use the spaCy English NLP models.

Intrinsic entity errors. To inject intrinsic entity errors into a summary S, we construct a dictionary of all unique entities appearing in the source document for S **only**, organized by entity label type. We then swap a random entity in the reference summary for a different entity of the same label type in the constructed dictionary.

Extrinsic entity errors. For extrinsic entity errors, we use the same dictionary construction for all unique entities appearing in **all** the corpus source documents. To change a random adjective, we use WordNet (Miller [1995]) to obtain the synsets for that adjective and swap the adjective for its antonym.

[3] `https://spacy.io/`

Pronoun entity errors. Pronoun errors are introduced with a preset list of commonly used pronouns. We randomly extract a pronoun set (e.g. she/her) from the text using the preset list and swap it with another random pronoun set (e.g. he/him).

Verb Negation. We use a rule-based system for verb negation based on verb tense, and predict tense based on the suffix and preceding words.

We note that injecting a certain level of error into a summary will have varying effects depending on the average length of summaries for a corpus. We use the same methodology for each corpus to maintain consistency, but future work may explore length-controlled error injection based on the objectives of the evaluation.

Natural Error Setting

In order to observe how metrics perform on machine-generated summaries, we generate summaries from fine-tuned T5 encoder-decoder summarization models (Raffel et al. [2020]) that was pretrained on news summarization data. We generate summary text using either beam search or sample-based decoding strategies. We then annotate the generated summaries for fine-grained factual errors to create a hand-curated factual consistency diagnostic dataset.

For human annotation of factual consistency in summaries, we show the source document, reference summary and a candidate summary that should be assessed for factuality. We then ask a factuality question with three choices:

- Yes (i.e. the summary is factual)
- No (i.e. the summary contains factual inconsistencies)
- Not Sure (i.e. the summary is too incoherent to judge)

If a summary is judged to be factually incorrect, annotators are allowed to select the number and type of errors they observe using a predefined list of factual errors. For less obvious cases of factual inconsis-

tency (for example when summaries contain locations or political figures that require regional background knowledge), we check factuality using external knowledge bases to ensure correctness of annotation. We also adhere to a strict binary notion of factuality in deciding cases where summaries are imprecise but ambiguous in terms of correctness, opting to label these summaries as factually inaccurate. If summaries are completely incoherent, we treat these summaries as having the highest level of factual inconsistency.

We validated the effectiveness of the setup by computing inter-annotator agreement of in-house expert annotators for 30 XSUM summaries. We achieve “fair" agreement of Krippendorff's α = 0.32 with 3 annotators and “moderate" agreement of α = 0.44 with 2 annotators (Landis and Koch [1977]; Ageeva et al. [2015]). The remaining annotations are done by one in-house expert annotator.

4.5 Meta-Analysis of Factuality Metrics

4.5.1 Controlled Data Experiments

We provide the results of the sensitivity analysis over our controlled data on the XSUM domain in Table 4.2, on CNNDM in Table 4.3 and on SAMSUM in Table 4.4. Our analysis reveals that QA metrics, ROUGE-(2/3) and BERTScore generally perform well at evaluating factuality. In contrast, ROUGE-(1/L) are frequently invalid as factuality metrics (Tables 4.2 and 4.3), and the performance of Cloze metrics varies across domains (BLANC-Tune is invalid on XSUM, but does fairly well on other domains). Also, performance of metrics tends to be much lower on news domains when we consider non-entity-based errors with the exception of QA-based metrics, ROUGE-(2/3) and BERTScore, indicating that while factuality and standard metrics are fairly attuned to changes in factual consistency that relate to entity-based errors, they are less robust to other types of factual errors.

4.5.2 Comparison with Human Evaluation of Model Generations

We find that metrics displaying invalid behavior on controlled data (for instance assigning higher metric values to more factually inconsistent summaries on XSUM in Table 4.2) also display this invalid behavior in model generations (Table 4.5). This indicates that meta-evaluation with controlled data is effective as a diagnostic tool for finding weak factuality metrics, and follows our intuition that non-entity errors, while fre-

	CLOZE		QA			STANDARD and CONTEXTUAL				
	BLANC-Help	**BLANC-Tune**	**SummaQA-C**	**SummaQA-F1**	**FEQA**	**R-1**	**R-2**	**R-3**	**R-L**	**BERTScore**
Upper Bound	5.99	1.73	9.64	4.48	27.87	10.61	2.56	0.72	9.32	83.76
Level 1	5.73 / 5.98	1.74 / 1.71	9.44 / 9.44	3.80 / 4.31	23.20 / 26.94	10.49 / 10.76	2.54 / 2.56	0.70	9.22 / 9.42	83.53 / 83.56
Level 2	5.46 / 5.99	1.59 / 1.78	9.27 / 9.35	3.40 / 4.22	20.05 / 26.55	10.40 / 10.86	2.51 / 2.54	0.69 / 0.68	9.16 / 9.49	83.36 / 83.38
Level 3	5.30 / 5.97	1.58 / 1.76	9.16 / 9.23	3.13 / 4.14	15.81 / 26.06	10.33 / 10.92	2.49 / 2.52	0.69 / 0.67	9.10 / 9.55	83.21 / 83.26
Lower Bound	0.51	-0.14	1.28	0.26	1.18	5.44	0.39	0.01	4.94	80.08
Correlation	-0.99 / -0.61	-0.88 / **0.69**	-0.99 / **-1.00**	-0.99 / **-1.00**	**-1.00**	**-1.00** / **0.98**	-0.97 / **-1.00**	**-0.87** / **-1.00**	**-1.00** / **1.00**	**-1.00**
p-value	0.09 / 0.59	0.32 / 0.51	0.07 / 0.05*	0.07 / 0.03*	0.05* / 0.04*	0.03* / 0.10	0.16 / 0.05*	0.33 / 0.05*	<0.01** / 0.02*	0.02* / 0.06

Table 4.2: Results of simulated factual error data experiments (**XSUM**, average of 5 runs, **=significant for p $\leq$ .01, *=significant for p $\leq$ .05). **For cells with (·/·), results for entity errors are reported on the left, results for non-entity errors are reported on the right.** The details for the upper/lower bounds, *p*-value and correlation measures are explained in §4.3.1. For sensitivity to factual consistency and correlation w/ factuality levels, we highlight the best-performing and lowest-performing metrics in **green** and **red** respectively. For cases where metric values are invalid (e.g. the metric values increase as factuality decreases), we highlight in **purple**.

	CLOZE		QA			STANDARD and CONTEXTUAL				
	BLANC-Help	**BLANC-Tune**	**SummaQA-C**	**SummaQA-F1**	**FEQA**	**R-1**	**R-2**	**R-3**	**R-L**	**BERTScore**
Upper Bound	7.60	5.79	13.82	10.87	37.56	14.33	8.08	4.75	13.83	84.36
Level 1	7.29 / 7.50	5.56 / 5.69	13.30 / 13.53	9.58 / 10.63	33.35 / 36.64	14.11 / 14.37	7.78 / 7.91	4.51 / 4.57	13.60 / 13.84	84.13 / 84.20
Level 2	7.03 / 7.43	5.43 / 5.58	12.93 / 13.24	8.53 / 10.38	28.46 / 36.13	13.95 / 14.38	7.55 / 7.75	4.32 / 4.40	13.44 / 13.85	83.94 / 84.04
Level 3	6.72 / 7.38	5.23 / 5.53	12.54 / 13.04	7.54 / 10.26	25.12 / 35.63	13.82 / 14.38	7.35 / 7.62	4.14 / 4.27	13.29 / 13.85	83.77 / 83.90
Lower Bound	-0.67	-0.19	1.61	0.12	0.58	5.85	0.47	0.02	5.55	78.16
Correlation	**-1.00** / -0.99	**-0.99** / -0.97	**-1.00** / **-1.00**	**-1.00** / -0.98	**-0.99** / **-1.00**	**-1.00** / **0.96**	**-1.00**	**-1.00**	**-1.00** / **0.91**	**-1.00**
p-value	0.03* / 0.08	0.07 / 0.17	0.01** / 0.06	0.01** / 0.13	0.07 / <0.01**	0.04* / 0.17	0.02* / 0.04*	<0.01** / 0.04*	0.03* / 0.27	0.01** / 0.02*

Table 4.3: Results of simulated factual error data experiments (**CNNDM**, average of 5 runs). (See Table 4.2 caption for details.)

	CLOZE		QA			STANDARD and CONTEXTUAL				
	BLANC-Help	**BLANC-Tune**	**SummaQA-C**	**SummaQA-F1**	**FEQA**	**R-1**	**R-2**	**R-3**	**R-L**	**BERTScore**
Upper Bound	15.23	10.13	13.83	17.23	55.36	26.55	8.24	4.07	25.06	84.60
Level 1	13.97 / 15.03	9.00 / 9.47	13.48 / 13.52	15.00 / 16.71	45.31 / 54.25	25.31 / 26.18	7.85 / 7.86	3.84 / 3.73	23.91 / 24.69	84.42 / 84.38
Level 2	12.87 / 15.01	8.36 / 9.46	13.16 / 13.26	12.26 / 16.50	37.01 / 53.10	24.27 / 25.86	7.60 / 7.59	3.68 / 3.50	22.99 / 24.38	84.28 / 84.19
Level 3	12.02 / 14.93	7.74 / 9.36	12.99 / 13.21	10.12 / 16.24	29.62 / 52.34	23.23 / 25.58	7.32 / 7.36	3.48 / 3.35	22.01 / 24.12	84.13 / 84.07
Lower Bound	0.92	-0.53	7.86	0.10	0.55	5.33	0.23	0.01	5.09	80.79
Correlation	**-1.00** / -0.96	**-1.00** / **-0.91**	**-0.99** / -0.94	**-1.00**	**-1.00** / -0.99	**-1.00**	**-1.00**	**-1.00** / -0.99	**-1.00**	**-1.00** /-0.99
p-value	0.05* / 0.18	0.01** / 0.28	0.11 / 0.23	0.05*	0.02* / 0.07	<0.01** / 0.03*	0.03*	0.05* / 0.08	0.01** / 0.04*	0.01** / 0.07

Table 4.4: Results of simulated factual error data experiments (**SAMSUM**, average of 5 runs). (See Table 4.2 caption for details.)

quently produced by abstractive summarization models, are difficult for standard summarization metrics to identify. When considering better-performing factuality metrics identified by the controlled error analysis, we find that the controlled data analysis is generally able to identify better-performing metrics (SummaQA, ROUGE-(2/3) and BERTScore) for XSUM with the exception of FEQA (FEQA metric performs well on XSUM controlled analysis (Table 4.2), but only approaches this performance on SAMSUM when we consider human eval). The strong overall performance of ROUGE-3 is consistent with the findings of Fabbri et al. [2020] on CNNDM, our work confirms that this metric is more consistently correlated with factuality than other ROUGE variations across domains.

Metric	XSUM		SAMSUM	
	Corr (- ←)	***p*-value**	**Corr** (- ←)	***p*-value**
BLANC-Help	0.04	0.55	-0.01	0.82
BLANC-Tune	0.00	0.98	-0.03	0.64
SummaQA-C	-0.11	0.11	-0.09	0.18
SummaQA-F1	-0.12	0.07	-0.14	0.03*
FEQA	0.04	0.57	-0.03	0.69
R-1	0.07	0.19	0.01	0.82
R-2	-0.10	0.15	-0.03	0.59
R-3	-0.12	0.07	-0.09	0.18
R-L	0.07	0.13	0.01	0.83
BERTScore	-0.17	0.01**	0.03	0.64

Table 4.5: Correlation (Corr) for 250 annotated XSUM and 250 SAMSUM generated summaries with fine-grained labeling. The arrow next to "Corr" indicates the direction of a correct correlation.

4.6 Summary of Key Findings

Standard summarization metrics are not always valid measures of factuality. ROUGE-1 and ROUGE-L fail to accurately measure factual inconsistency across domains in our controlled analysis. The ROUGE-L results raise the question of context *relevance*. While ROUGE-L takes into account more context than other ROUGE variations, this context may not be relevant for assessing factuality. For example, swapping "decreased" for "increased" dramatically changes the meaning in the summary *"Scotland's renewable energy output increased by 45% in the first quarter of this year, compared with the same period last year."*, but ROUGE-L is not affected. Despite the frequent use of ROUGE-L as a more contextual measure, prior work has also noted that ROUGE-N outperforms ROUGE-L (Rankel et al. [2013]; Fabbri et al. [2020]).

Analysis on human annotated data is still necessary as an upper-bound on *meta*-evaluation quality. While BLANC-Help, FEQA metric and BERTScore values decrease with factual inconsistency on controlled data, the metrics may sometimes be positively correlated with factual inconsistency on generated data. This emphasizes the importance of a expert curated test set as part of the GO FIGURE meta evaluation for the most rigorous testing. **A question-answering objective is promising for measuring factual consistency across domains, but effectiveness depends on the question.** While QA metrics can perform well at measuring factual consistency of generated summaries, our *meta*-evaluation reveals this is dependent on the way in which questions are asked. While both QA metrics use SQuAD-based systems (Rajpurkar et al. [2016]), asking questions from the source rather than the summary is most robust across domains. This

opens the door to metrics based on more contextual QA like commonsense (Shwartz et al. [2020]).

Chapter 5

A Study of Question-Answering Robustness

The advent of large language models could revolutionize every area of our lives, including healthcare. Given the strong performance of technologies like GPT-* models, tech companies are already beginning to consider medical applications of machine-generated text.[1] This opens up the question of how all this machine-generated content will be evaluated in high-risk environments like hospitals. Many existing works focused on medical text rely on the same flawed metrics used for other forms of text generation (Zhang et al. [2018b]; van den Bercken et al. [2019]; Devaraj et al. [2021]; Delbrouck et al. [2022]). As highlighted by the last chapter, a promising direction in factuality evaluation is the alignment of evaluation with question-answering based objectives. It seems intuitive that an informative generated summary or continuation should be more useful for answering questions about the context than a false one. It is also more human-interpretable, in that all intermediate components (e.g. generated questions) used for evaluation are visible. However, our meta-evaluation shows the potential brittleness of the approach. Since proposed QA-based metrics primarily rely on SQuAD Rajpurkar et al. [2016], there is no reason to assume that they would work as well on diverse text (e.g. text with significant technical jargon like radiology reports). As an immediate exploration following from this work, we concretely show the high variance of the question-answering objective at assessing quality and factual correctness in medical text generation.

[1] https://www.forbes.com/sites/katiejennings/2023/05/01/microsoft-wants-to-automate-medical-notes-with-gpt-4--but-doctors-need-convincing/?sh=603320711d27

	Medical - Useful				Medical - Factual			
	Pears.		Spear.		Pears.		Spear.	
	ρ	p-val	r	p-val	ρ	p-val	r	p-val
- F1	0.05	0.63	0.07	0.48	0.01	0.93	-0.12	0.25
- Prob	0.10	0.30	0.15	0.13	0.01	0.96	0.03	0.78

Table 5.1: Pearson and Spearman Rank correlation of QA metric scores across different question types for GPT-3 generated medical report summaries.

5.1 Problem Setup

As described in Chapter 5, question-answering (QA) metrics measure how well a generated document (e.g. a summary) S_i helps in answering questions about a source document D_i. We use the same Bert-based evaluation framework from (Scialom et al. [2019]).

5.1.1 Testing Data

Here we use 100 articles from TrailStreamer (Marshall et al. [2020]) describing randomized clinical trials for healthcare interventions. These articles were originally collected by Shaib et al. [2023], who use GPT-3[2] to generate a summary for each medical document. They ask 3 domain experts from Upwork[3] to annotate the GPT-3 generated summaries based on coherence, faithfulness and usefulness. We use their ratings for factual errors (a 3-pt scale ranging from no error to major error) and usefulness (a 4-pt scale ranging from strongly disagree to strongly agree) to compare effectiveness of QA metric variations.

5.2 Results

From Table 5.1, we can see that while prior work found there to be significant agreement between QA metric scores and human ratings of relevance for newswire generated summaries,[4] there is no evidence for this correlation on medical text. These findings should be treated as a cautionary tale to avoid over-confidence in performance of generation metrics on domain-specific text. We hope this case study and the preceding meta-evaluation will motivate innovative solutions that prevent harms from overgeneralization.

[2]They specifically use the "text-davinci-003" version of the model.

[3]`https://www.upwork.com/`

[4]Scialom et al. [2019] reports a statistically significant correlation of Spearman $r = 0.31$.

Part II

Harm

Chapter 6

Controlled Generation for Harm Mitigation

Recent work has raised serious concerns about potential harms introduced by large language models (Bender et al. [2021]; Kumar et al. [2022]). These range from the risk of natural toxic degeneration due to models overweighting undesirable language patterns learned from scraped web data (Gehman et al. [2020]), to purposeful misuse (e.g. through data poisoning) to manipulate human behavior (Wallace et al. [2021]). This highlights the double-edged sword of remarkable progress in coherence of machine-generated language due to self-supervised learning - neural language models can now emulate a diverse range of human communication, including extremist language (McGuffie and Newhouse [2020]) and hard-to-detect implicit toxicity (ElSherief et al. [2021]; Hartvigsen et al. [2022]; Gabriel et al. [2022]). Approaches have been proposed that consider mitigating harmful behavior through *controlled generation* (Hu et al. [2017]), for example with inference time interventions (Dathathri et al. [2020]; Sheng et al. [2020]; Liu et al. [2021a]). A counter-intuitive approach is to specifically encourage or take advantage of undesirable behavior in order to prevent harms, often through contrastive learning (Adolphs et al. [2022]) or using machine-generated harmful language as augmentation data (Wullach et al. [2021]). While these methods can be effective, there are legitimate ethical concerns around such use. Zellers et al. [2019] introduced a release strategy to address dual-use technologies like neural disinformation generators used for detecting false language. The core point underlying this strategy is that public safety demands open access to defenses that match the computational power of technology available to attackers. We argue that while limited and dangerous if misused, the vast potential of large language models for improving social understanding and safety (Ziems

et al. [2023]) should not be ignored. To handle emerging types of attacks like bots designed to manipulate political processes (Woolley [2020]), we need a concrete understanding of the threat landscape and a robust means to directly counter such attacks. Other works within the area of AI policy (Learned-Miller et al. [2020]; OSTP [2022]) hone in on use-case, advocating for an approach to technology regulation that fully considers the context and intention with which such technologies are deployed.

While the purpose of our work is to curate diverse and effective hate speech detection resources, our methods encourage a large language model to make its generation *more* toxic. This poses a potential misuse case where bad actors exploit these methods for nefarious purposes like spreading machine-generated hate speech. Still, ignoring this possibility does not make it go away and our work introduces an opportunity for the community to push back against harm towards minority groups. Our ultimate aim is to shift power dynamics to targets of oppression ... We see a path forward in which tools and techniques like those presented in this work are paired with human expertise and well-informed policy & regulation in bringing scalable and reliable solutions to practice. We acknowledge and encourage the critical role the NLP research community is poised to play in this inter-disciplinary effort. - From Hartvigsen et al. [2022]

The following chapter describes the implementation of this approach and a related work (Gabriel et al. [2022]), which use samples generated from a pretrained language model to close gaps in existing hate speech benchmarks. In the next few sections, we provide background for the hate speech and toxicity detection tasks. We also describe the basic adversarial learning setup.

6.1 How should we define harm?

Human perception of harm is complex. A seemingly innocuous statement to one reader can be glaringly offensive to another. A single reader's perception of a text's harmfulness may shift depending on who they imagine as being the author or the audience. Recent works have explored the role socio-cultural factors play, and how annotator agreement in this setting may indicate favorable diversity rather than a design flaw (Davani et al. [2021]; Sap et al. [2022b]). We summarize common definitions for harmful language below:

- According to the UN,[1] ***hate speech*** is "*any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor*." This type of language is used to dehumanize and dominate victims, sometimes villainizing them, and includes use of ethnic slurs. Waseem and Hovy [2016] propose a theoretically grounded decision list for labeling social media text as hate speech.

- ***Toxicity, offensiveness*** and ***abuse*** are more broadly defined (Davidson et al. [2017]; Jurgens et al. [2019]). While they include hate speech, they also cover harmful language which is not group-directed (e.g. personal insults that are not based on identity).

- ***Microaggressions*** refer to a subset of implicit toxicity that "*subtly or often unconsciously expresses a prejudiced attitude toward a member of a marginalized group such as a racial minority*" (Breitfeller et al. [2019]).

- ***Stereotyping*** is the construction of generalized social beliefs about people based on their identity group. All types of stereotyping, even positive stereotyping (Czopp et al. [2015]), can be harmful due to the manner in which they overgeneralize, propagate false beliefs and reinforce social biases (Nangia et al. [2020]).

- We refer to toxic language written with a humorous or ironic intent as ***cruel humor***. While the original intent of this type of language may be benign, the content can still have a negative effect on readers.

- ***Lewd content*** and ***profanity*** refers to language of an inappropriately sexual nature and swearing respectively. Offensiveness depends heavily on the social context, including the maturity of the target audience and cultural norms.

6.2 Adversarial Learning for Harm Mitigation and Beyond

We can conceptualize adversarial machine learning as a game between a defender choosing a learning algorithm H and an adversary choosing an attack algorithm A (Huang et al. [2011]). The objective of the

[1] `https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech`

defender is to select H to maximize prediction performance. Simultaneously, the adversary uses A to determine inputs to H that will minimize performance by fooling the algorithm, known as “adversarial examples." Attacks are perpetrated by an adversary who may have full knowledge of the learning algorithm or model, as in the *white-box* attack setting, or is completely unaware of the model, as in the *black-box* attack setting (Sablayrolles et al. [2019]). *Adversarial robustness* measures the susceptibility of a model to these attacks (Salman et al. [2020]). While performance of machine learning algorithms on adversarial examples were initially explored in the computer vision domain (Szegedy et al. [2013]; Goodfellow et al. [2015]; Hendrycks et al. [2019]) follow-up works have uncovered their usefulness for improving robustness and understanding of language models (Miyato et al. [2017]; Zhao et al. [2018]; Ebrahimi et al. [2018]; Jin et al. [2020]). A less explored area is adversarial learning for improving toxic language and hate speech classifiers (Rusert et al. [2022]). In the next chapter, we propose algorithms for significantly improving the robustness of toxic language classifiers.

Chapter 7

LLM-Driven Toxicity Detection

Transformer models have gained prominence in NLP research due to their powerful performances on leaderboards. However, numerous studies have shown these neural models are brittle, frequently taking shortcuts to reach decisions rather than reasoning about the underlying semantics correctly (Geirhos et al. [2020]; Bras et al. [2020]) or failing when exposed to adversarial perturbations of inputs (e.g., Goodfellow et al. [2015]; Szegedy et al. [2013]; Jia and Liang [2017]; Glockner et al. [2018]; Dinan et al. [2019]). Due to the opaque nature of neural modeling, methods for adversarial example generation may also steer algorithms towards generating unlikely examples that exhibit unrealistic properties (Zhao et al. [2018]). In this chapter, we explore how to control large language models in order to use them as adversarial example generators for hate speech and toxicity classifiers. We pose the question, *"what does it really mean for an adversarial attack to be effective and can naturalistic adversaries match artificial ones?"* We argue that effectiveness should be dependent not only on attacking accuracy, but on usefulness of adversaries for improving robustness under realistic conditions (e.g identifying social biases learned by neural models, Buolamwini and Gebru [2018]; Sheng et al. [2019]; Stanovsky et al. [2019]; Sap et al. [2019b]; Ross et al. [2021]).

In §7.2, we propose a framework NaturalAdversaries[1] for generating convincingly naturalistic adversaries. We first approximate the behavior of a given classifier decision function $F_c(x)$ and then train a generative model $F_g(x)$ to mimic this behavior. We condition generative models on influential tokens extracted using black box or white box explainability methods (Ribeiro et al. [2016]; Sundararajan et al. [2017]), and

[1]Code and data can be found here: `https://github.com/skgabriel/NaturalAdversaries`.

a desired label (e.g. "*entailment*" or "*contradiction*"), to produce new examples that match a distribution learned from $F_c(x)$ through sampling.

In §7.4, we describe the creation of ToxiGen,[2] a large-scale and machine-generated dataset of 274k toxic and benign statements about 13 minority groups. We develop a demonstration-based prompting framework (§7.4.1) and an adversarial classifier-in-the-loop decoding method (§7.4.2) to generate subtly toxic and benign text with a massive pretrained language model. Controlling machine generation in this way allows ToxiGen to cover implicitly toxic text at a larger scale, and about more demographic groups, than previous resources of human-written text.

7.1 Motivation: Limitations of Hate Speech Detection

Recent work (Blodgett et al. [2021]) identifies concerning issues in both conceptualization and operationalization phases of dataset construction for measuring harmful behaviors like stereotyping. For example, existing datasets may conflate types of potentially harmful language, e.g. not distinguishing between overtly offensive language like hate speech and positive stereotyping. They also contain unnatural examples (e.g. "*While little black / white Drew watched, his father went off to prison*" from CrowS-Pairs (Nangia et al. [2020])) which are less beneficial for mitigating harmful behavior in real-world settings, and coverage across identity groups is limited. Furthermore, datasets are imbalanced. Minority mentions are more often the targets of social biases and toxicity (Hudson [2017]). As such, minority mentions often co-occur with toxicity labels in datasets scraped from online platforms (Dixon et al. [2018]). For example, over 93% of mentions of Jewish folk in the Social Bias Frames corpus (Sap et al. [2020]) are toxic (Wiegand et al. [2021]). Models trained on such data can exploit these spurious minority-toxicity correlations instead of considering the deeper semantics of text (Zhou et al. [2021]). Importantly, the spurious correlations are also learned by large language models, which are known to produce stereotypical, biased, or toxic content when prompted with minority mentions (Sheng et al. [2019]). Given that the main mitigation approach to prevent large language models from generating toxic language is to train new classifiers to detect such language, these classifiers *also* learn the spurious correlations and start blocking most language referencing minority groups.

[2]Code and data can be found here: `https://github.com/microsoft/TOXIGEN`.

7.2 NaturalAdversaries

In this section, we introduce an adversarial generation framework for inducing large language models to generate naturalistic examples that fool hate speech classifiers. Here we focus on uncovering potentially unknown lexical biases (§7.4 covers mitigation of known biases).

7.2.1 Defining Naturalness

We first define what it means for machine-generated adversaries to have the quality of "naturalness." Prior work on text generation has defined this property in terms of linguistic competence of text (Novikova et al. [2018]; Lau et al. [2020]), as well as enumerating undesirable characteristics that lower perceived naturalness like self-contradiction (Dou et al. [2021]). In our work, we ask human evaluators to judge naturalness in terms of whether generated text fragments are *coherent*, *well-formed* and *likely to be human-written*.

7.2.2 Description of NaturalAdversaries

Our overall framework consists of two stages - (1) a probing stage where we identify the influential (high attribution) tokens and (2) an adversarial generative stage where we generate unseen challenging examples by conditioning on the extracted tokens and a reversed label (§7.2.3). We focus on two types of explainability methods as a means of summarizing model behavior through sampling - (1) LIME local linear explanation models (Ribeiro et al. [2016]) and (2) gradient attribution scores (Sundararajan et al. [2017]). Using these methods as part of our proposed method is advantageous since it doesn't require curation of cherry-picked examples to probe model behavior, and is agnostic to the specific internal structure of a given classifier. Given a sequence of text tokens $S = [s_0, s_1, \ldots, s_i, s_{i+1}, \ldots]$ with classifier label $\hat{y}$, each of these methods define a scoring function $F_{attr}(s_i)$ which we use for assigning attribution scores to each token s_i which represents its overall contribution to the classifier's decision. We separate these approaches based on whether F_{attr} is conditioned on the model parameters or not. If it is, this demonstrates a white-box attack that can adapt to the vulnerabilities of a specific classifier given complete access to the model (e.g., using gradient attribution scores). In the black-box setting (e.g. using LIME attribution scores), the underlying architecture and parameters are not known, and only sampled predictions are used to approximate the model behavior.

Dataset	Taxonomy	Example	Classifier(s)
DynaHate (Vidgen et al. [2021])	hate / nothate	I say I like women, but I don't	RoBERTa, BERT
ANLI (Nie et al. [2020])	contradiction / neutral / entailment	**P**: P-17 is a mixed use skyscraper proposed for construction in Dubai... The design is for a 379 m tall building, comprising 78 floors. **H**: P-17 is designed to have 78 floors and be over 500 meters tall.	DeBERTa, BERT

Table 7.1: Description of considered datasets. For ANLI, each example comes with a premise (**P**) providing context and a hypothesis statement (**H**).

7.2.3 Adversarial Generation

Given a generative autoregressive model F_g and training set D_1 with triples $(S, y, F_{attr}(S))$, we construct the following input sequence

$$x = [attr, z, label, y', text, S, eos] \tag{7.1}$$

where z is a sequence of influential tokens sampled from S using the attribution weights defined by $F_{attr}(S)$, $attr$ is a special token indicating the start of this sequence, y' is the desired classification label, $label$ and $text$ are special tokens indicating the start of y' and S respectively, and eos is a special token indicating where the full input sequence ends. At training time, $y' = y$ as the generated model is trained to mimic the behavior of the classifier model and generate examples with a given label y' based on the classifier's observed behavior. At decoding time, we encourage adversarial behavior by reversing the label (e.g. setting $y' = 1 - y$). The model is prompted using only the influential tokens z and y'. For example, given a natural language inference (NLI) premise and hypothesis pair (*"It was sunny outside"*, *"it was too dark to see anything outside"*) where the gold label is contradiction, at training time we use (y'="contradiction";z="influentialWord1", "influentialWord2", "influentialWord3", S=*"It was sunny outside. It was too dark to see anything outside."*). At decoding time we would use (y'="entailment";z="influentialWord1", "influentialWord2", "influentialWord3") and predict S.

We minimize cross-entropy loss during training time:

$$\mathcal{L}_{CE} = -\frac{1}{|S|}\sum_{i=1}^{|S|} logP(S_i|S_1, ..., S_{i-1}). \tag{7.2}$$

7.2.4 Experimental Setup

In this section we first introduce the domains we test on (§3.2.1) and then methods used for baseline comparison (§7.2.5). All adversarial generators are based on the GPT-2 124M parameter model. We describe evaluation setups (§7.2.6), as well as out-of-distribution evaluation (§7.2.6).

An advantage of proposed generative method is that we can automatically extend human-in-the-loop adversarial generation methods like Adversarial NLI (ANLI, Nie et al. [2020]), which are costly and time-consuming to curate. Given this motivation, we focus on particularly challenging human-in-the-loop examples rather than cases that are already well solved by existing benchmarks. To study effectiveness of the proposed approach, we conducted experiments on the hate speech detection (DynaHate, Vidgen et al. [2021]) and natural language inference (NLI). For DynaHate we use a RoBERTa classifier trained on tweets (Founta et al. [2018]; Zhou et al. [2021]). We test generalization across both model architectures and (non-adversarial) data domains using a BERT model (Devlin et al. [2019]) trained on the HateXplain dataset (Mathew et al. [2020]). For NLI we use DeBERTa (He et al. [2021]) trained on MNLI (Williams et al. [2018]). We test generalization using BERT trained on the QNLI dataset (Wang et al. [2018]).

7.2.5 Baselines

For automatic adversarial example construction, we compare against several common adversarial example generation approaches which are designed for either **black-box** (model-agnostic) or **white-box** (model-dependent) attacks. Baselines are implemented using TextAttack (Morris et al. [2020]).

Black-Box Baselines We use the TextFooler (Jin et al. [2020]) algorithm for generating coherent adversaries by replacing high-importance words in original examples with words that preserve semantic similarity.

White-Box Baselines Since our approach has an advantage over other baselines in the white-box setting of utilizing knowledge about model parameters, we also compare against a word-level version of the widely

used HotFlip gradient-based approach (Ebrahimi et al. [2018]).

7.2.6 Evaluation Metrics

Human Evaluation

To compare effectiveness of automatic methods, adversarial examples are manually validated to determine the true label. We also assess *naturalness* of examples, i.e. whether they are perceived as realistic examples that could be written by humans. We use 156 crowd-source workers from Amazon Mechanical Turk (MTurk) with prior experience validating hate speech (Sap et al. [2020]) and 79 workers with experience validating NLI data (Liu et al. [2022]). We sample 150 examples using each approach (some approaches may impose constraints that are unsatisfied by all candidate transformed sentences, we also filter out examples that are already adversarial to avoid conflating adversarialness of original examples with effects of the transformation). Each example is judged by 3 different workers. For hate speech, we classify an example as toxic if at least one annotator considers it so. We achieve moderate inter-annotator agreement of Fleiss' $\kappa = .51$ for hate speech and $\kappa = .52$ for NLI.

Out-of-Distribution Performance

Here we frame domain adaptation as a few-shot learning problem, where the adversarial evaluation set represents training examples from outside the seen domain of the classifier. To test out-of-distribution (OOD) performance on hate speech data, we use the HateCheck test suite (Röttger et al. [2021]), which consists of test cases for 29 model functionalities relating to real-world concerns of stakeholders. For NLI we check OOD performance on the SNLI-Hard dataset (Gururangan et al. [2018]), which assesses common model vulnerabilities.

7.3 Results

We discuss results for the TextFooler (TF) and HotFlip (HF) baselines along with our two model variations (NA-LIME and NA-IG).

Dataset	Model	Natural$_H$ (%)	Adv1 (%)	Adv2 (%)	HateCheck (F1)
DynaHate	Original	-	-	-	55.01
	TF	53	**69**	**59**	55.59
	HF	27	<u>30</u>	<u>55</u>	<u>55.92</u>
	NA-LIME	<u>67</u>	<u>30</u>	<u>55</u>	55.90
	NA-IG	**73**	21	36	**56.69**
Dataset	Model	Natural$_H$ (%)	Adv1 (%)	Adv2 (%)	SNLI-Hard (F1)
ANLI	Original	-	-	-	76.95
	TF	57	**57**	**46**	76.82
	HF	64	<u>33</u>	38	**76.98**
	NA-LIME	<u>73</u>	31	<u>43</u>	**76.98**
	NA-IG	**89**	27	42	<u>76.97</u>

Table 7.2: Human evaluation (Natural$_H$) of naturalness, along with adversarial performance against the original target classifier $Adv1$ and an unseen classifier $Adv2$. In the last topright column we show macro-averaged F1 performance on HateCheck (Röttger et al. [2021]) after finetuning RoBERTa on 150 adversarial examples, compared to the original performance. We conduct a similar experiment for NLI using the SNLI-Hard evaluation set (Gururangan et al. [2018]) with results in the last bottomright column. We bold the best-performing model and underline the second best model.

Quality and effectiveness of adversarial generations. From Table 7.2, we can see that examples generated using NaturalAdversaries are perceived as more natural across domains (20% more for hate speech and 25% more for NLI). While attacking accuracy is generally lower than artificial adversaries, we also find that our black-box approach generalizes well to classifiers other than the original target model.

OOD performance. Although NLI model performance is relatively unaffected by finetuning, when we assess the RoBERTa classifier using HateCheck, we find that the target model exhibits concerning vulnerabilities. Finetuning generally improves performance, though our NA-IG model leads to the most improvement (1.68 F1 over base performance). Given the small size of our evaluation set (150 examples), this indicates naturalistic adversarial examples may address classifier weaknesses, with minimal need for manual annotation.

7.4 Toxigen

To address gaps in existing hate speech datasets, we use demonstration-based prompting for GPT-3, encouraging a text generator to produce both toxic and benign sentences that mention minority groups without using explicit language. We introduce a classifier-in-the-loop decoding method based on constrained beam search, ALICE,[3] which, along with samples generated without ALICE, contributes to generating a challenging corpus of over 274k toxic and harmless statements relating to 13 identity groups.

7.4.1 Prompting with Demonstrations

Prompts are text fragments passed into language models that can encourage certain behaviors (Brown et al. [2020]). However, designing prompts is notoriously challenging (Liu et al. [2021b]). While there are several approaches for prompting pretrained LLMs, a recent and promising direction is *demonstration-based prompting* (Gao et al. [2021b]; Mishra et al. [2021]). Here, example statements are passed to GPT-3, encouraging it to produce a similar, but distinct, statement. For example, to encourage to produce *benign* statements about *Asian* people, we might use the following examples: (1) *many impactful people are of asian descent*, (2) *there are hundreds of asian cultures*. To engineer prompts that lead to high-quality, group-mentioning toxic and benign statements at scale, we first gather and curate sets of examples. For benign prompts, we encourage realistic text generation and include diverse voices by collecting benign sentences from blog posts and news articles that mention a group. To encourage implicit toxicity from a LLM, we find examples of human-written sentences with implicit toxicity towards each group from hate forums (de Gibert et al. [2018]) and Reddit (Breitfeller et al. [2019]). In total, we create 26 sets of prompts, with two (benign and toxic) per identity group.

7.4.2 Adversarial Decoding

Demonstration-based prompting alone consistently produces toxic and benign statements about minority groups, but there is no guarantee that these statements will be challenging to existing toxicity detectors. Therefore, we also develop ALICE, a variant of constrained beam search (CBS) (Anderson et al. [2017]; Hokamp and Liu [2017]; Holtzman et al. [2018]; Lu et al. [2021]) during decoding that generates statements

[3] Adversarial Language Imitation with Constrained Exemplars

that are adversarial to a given pretrained toxicity classifier.

ALICE creates an adversarial game between a pretrained language model (PLM) and a toxicity classifier (CLF) during constrained beam search decoding. In many CBS settings, constraints are added during beam search decoding to force the model to either include or exclude a specific word or group of words in the output (Anderson et al. [2017]; Hokamp and Liu [2017]; Lu et al. [2021]). With ALICE, we instead want to enforce *soft* constraints on the probabilities coming from a given toxicity classifier CLF during beam search:

$$s(w_{i+1}) = \lambda_L \log p_{\text{LM}}(w_{i+1}|w_{0:i}) + \lambda_C \log p_{\text{CLF}}(w_{0:i+1})$$

Here, λ_L and λ_C denote hyperparameters that determine the respective contribution of the language model and classifier to the decoding scoring function. By using this weighted combination, we can steer generations towards a higher or lower probability of toxicity without sacrificing coherence enforced by the language model. To create examples that challenge existing toxicity classifiers, we use two adversarial setups:

- **False negatives**: We use *toxic* prompts to encourage the language model to generate toxic outputs, then maximize the classifier's probability of the *benign* class during beam search.
- **False positives**: We use *benign* prompts to encourage the language model to generate non-toxic outputs, then maximize the probability of the *toxic* class during beam search.

In the first approach, we are also able to detoxify model outputs when the classifier successfully steers the generations towards non-toxic language.

7.4.3 Experimental Setup

To verify the labels for adversarial examples and learn about harms conveyed by GPT-3 generated text, we use Amazon Mechanical Turk crowdworkers to obtain detailed annotations.

Human Validation Design

For each generated statement, we ask the annotators various questions, described below, that take into account multiple dimensions of how toxic machine-generated language presents a potential harm to readers.

Perceived hatefulness with respect to human or AI-authored text. We first ask annotators to guess whether a statement is human- or AI-generated (HUMANORAI). Then, we ask whether the statement would be harmful to anyone if an AI system wrote it (HARMFULIFAI), as well as if a human wrote it (HARMFULIFHUMAN); we hypothesize that readers may have different standards for machine-generated text than human-written text. For all questions measuring harmfulness of text, we consider potential harm on a 1-5 scale with 1 being clearly benign and 5 indicating very offensive or abusive text.

Perceived intent of the writer. We ask readers whether statements were likely intended to be harmful (HARMFULINTENT), since some biased statements can be positively intended (Glick and Fiske [1996]). Additionally, we ask if the statement exhibits a positive stereotype (POSSTEREO), which is also harmful (Cheryan and Bodenhausen [2000]).

Detailed harm explanations. To better understand how harm may be perpetrated against the minority group, we ask readers in-depth questions about a text fragment's content, following (Sap et al. [2020]; Olteanu et al. [2018]). We ask whether or not the statement is lewd or sexual (LEWD), whether and how it references the targeted group or other groups (WHICHGROUP, GROUPFRAMING), and whether it claims to be factual (FACTOROPINION).

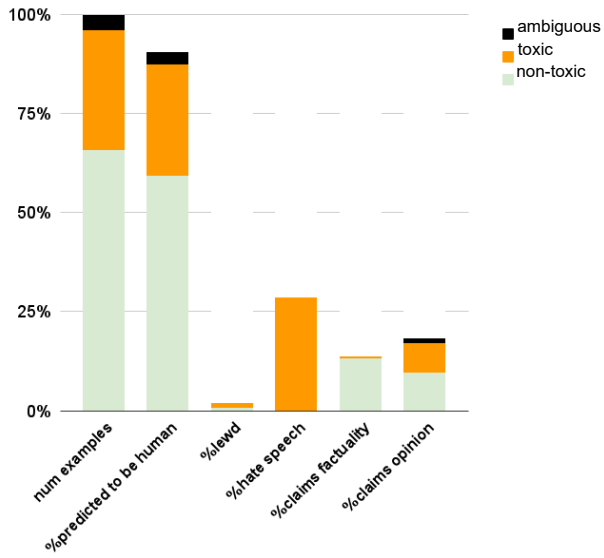

Figure 7.1: Summary statistics for the human annotations on the evaluation set. Each statistic that the annotators are asked to evaluate is shown along the x-axis, while the y-axis gives the percentage of examples per annotated class (non-toxic, toxic, ambiguous).

Validation Set

Data and Setup. We selected 792 statements from Toxigen to include in our test set, such that no training statement had cosine similarity above 0.7 with any test statement. Each test statement was then rated by 3 annotators from a pool of 156 prequalified annotators from Mechanical Turk.

Text	HumanOrAI (Predicted/True)	AvgScore (AI/Human)	Intent	groupFraming
I started wearing a head scarf at the age of 12 because my mom told me that it was my duty as a woman to cover my hair	Human/AI (ALICE)	1.3/1.3	Benign	indirect reference, moral judgement
if the discussion is for the sake of the greater black community rather than for you, quit trying to hypercentrally control it	Human/AI (top-k)	2.7/2.7	Unsure	direct reference, suggests solution, moral judgement
we do not need to mention trans people in an article about lgbtq+ issues	Human/AI (top-k)	4.0/4.3	Harm	direct reference, suggests solution
the mentally disabled should be able to live in their communities and not be forced into institutions.	Human/AI (ALICE)	1.0/1.0	Benign	direct reference, suggests solution, other framing

Table 7.3: Example responses from human evaluation where machine-generated text fools annotators into thinking the writer is human. Average toxicity scores are on a 1-5 scale (1 being benign and 5 being clearly offensive), and are averaged across annotator responses. We report scores for the case where annotators assume the writer/speaker is AI and the writer/speaker is human respectively.

Inter-annotator agreement. To investigate the quality of our annotations, we compute agreement on toxicity ratings.[4] We find that annotators agreed moderately and are higher than or equal rates to prior work on hate speech annotation (Sap et al. [2020]), with a Fleiss' κ=0.46 (Fleiss [1971]) and Krippendorff's α=0.64 (Krippendorff [1980]). In 55.17% of cases, all 3 annotators agree, while a majority ($\geq$2/3) agree for 93.4%.

7.4.4 Results and Summary of Key Findings

First, **we find that our machine-generated statements are largely indistinguishable from human-written statements**. For example—see Table 7.3—human annotators often predict that our text is generated by a human. In fact, on average 90.5% of machine-generated examples are thought to be human-written by a

[4]Specifically, we take the max of the HARMFULIFAI and HARMFULIFHUMAN scores and map it into three classes (scores $<$3: "non-toxic", =3: "ambiguous", $>$3: "toxic").

	Test Data	Finetune Data			
		None	ALICE	top-k	ALICE + top-k
HateBERT	SBF_{test}	0.60	0.66	0.65	**0.71**
	IHC	0.60	0.60	0.61	**0.67**
	DYNAHATE	0.47	0.54	0.59	**0.66**
	TOXIGEN-VAL	0.57	0.93	0.88	**0.96**
RoBERTa	SBF_{test}	0.65	**0.70**	0.67	**0.70**
	IHC	0.57	0.64	0.63	**0.66**
	DYNAHATE	0.49	0.51	0.50	**0.54**
	TOXIGEN-VAL	0.57	0.87	0.85	**0.93**

Table 7.4: AUC for HateBert and RoBERTa both zero-shot and fine-tuned on 3 versions of our dataset: ALICE only, top-k only, and both combined. Since there are fewer ALICE samples than top-k, we downsample top-k for fair comparison via equal-sized datasets. ALICE + top-k combines these two datasets.Each model is evaluated on three external human-written datasets and the human-validated portion of Toxigen. Bolding denotes the best performance. In the zero-shot setting (first column) ALICE creates more challenging evaluation samples by attacking HateBERT and RoBERTa.

majority of annotators, as shown in Figure 7.1. We also note that harmful text confuses readers slightly more than non-harmful text: 92.9% of toxic examples are mislabeled as human-written compared to 90.2% for non-toxic. Most toxic examples are also hate speech (94.56%). While opinions are common in both toxic and non-toxic examples, most fact-claiming text is non-toxic.

Interestingly, **there is little difference in toxicity when we account for whether annotators perceive scores as written by humans or AI**. This finding indicates that our machine-generated text is perceived as similarly harmful to human text. We also find that the most common framing tactic is "moral judgement", or questioning the morality of an identity group, which has been linked to toxicity by prior work (Hoover et al. [2019]).

To further showcase the usefulness of Toxigen, we investigate how it can enhance classifiers' abilities to detect human-written and machine-generated implicit toxic language. We fine-tune the widely-used HateBERT (Caselli et al. [2021]) and ToxDectRoBERTa (Zhou et al. [2021]) models on the training portion of Toxigen, using the prompt labels as proxies for a true toxicity label. Then, we compare the performance of the out-of-the-box models to those fine-tuned on Toxigen on three publicly available human-written datasets (IMPLICITHATECORPUS (ElSherief et al. [2021]), the SOCIALBIASFRAMES test set (Sap et al. [2020]),

and DYNAHATE (Vidgen et al. [2021])) as well as the evaluation portion of our machine-generated dataset. To ablate the contribution of each decoding method, we also split Toxigen into equal numbers of ALICE-generated and top-k-generated examples.

Our results—see Table 7.4—show that fine-tuning HateBERT and ToxDectRoBERTa on Toxigen improves performance across all datasets. Toxigen can be used to improve existing classifiers, helping them better tackle the challenging human-generated implicit toxicity detection task. Fine-tuned HateBERT performs strongly on the Toxigen eval set, demonstrating that our data can successfully help guard against machine-generated toxicity.

Part III

A Vision for Community-driven NLP

Chapter 8

Unified Factuality and Harm Detection

In the earlier chapters, we highlighted how machine learning algorithms fail by interpreting language out-of-context. We also showed through studies like Hartvigsen et al. [2022] that despite their limited exposure to the human experience, they can be uncanny mimics of human behavior. In modern day Artificial Intelligence research, we cannot separate machines from people. They inevitably memorize and encode our disparate values (Birhane et al. [2022]; Jiang et al. [2022]), providing a sometimes convoluted lens on a polarized society. We argue that one of the most critical questions for the future is not how to impose a concept of neutrality on machines, which may be impossible. Instead, a critical question is how to ensure their interpretation of our world considers the multi-dimensional aspects of human culture. In Chapter 3, we showed how we can begin to implement this by modeling human behaviors with pragmatic frames of intent and impact. However, we need diverse community feedback for such approaches to be fully effective. This calls for a grassroots approach to NLP research, in which domain experts and laypeople work collaboratively to ensure responsible development of AI systems. In the next section, we briefly describe a conceptual framework for community collaboration to mitigate misinformation.

8.1 Continual Learning for Misinformation Detection

Top news stories rapidly go out-of-date and are replaced, e.g. during political election cycles. A recent study of Google Trends[1] in 2018 found that popular news stories tend to stay relevant for a lifespan of only

[1] https://www.newslifespan.com/

7 days. The actual observed behavior of the general public seems incongruous with the current paradigm of automated fact-checking, which relies on static resources (e.g. extracted knowledge bases, pretrained language models) to ensure well-informed public discourse (Schuster et al. [2021]). Static knowledge bases like evidence retrieved from Wikipedia archives (Thorne et al. [2018]; Gupta et al. [2022]) are not guaranteed to remain relevant as media narratives shift over time. While pretrained language models (PLMs) have been conceptualized as knowledge bases (Petroni et al. [2019]; West et al. [2022]; AlKhamissi et al. [2022]), Chapter 4 showed that the veracity of information stored by neural language models may be too unreliable for fact-checking without filtering. This information is also brittle to an evolving world and may not generalize well to reasoning about current events (Bender et al. [2021]). We propose *Unified Misinformation Detection (UMD)* as a solution:

Algorithmic contribution. UMD would consist of a transformer-based misinformation detection system with (1) a unified fact verification model M_{FC} that predicts the veracity and potential harm (e.g. toxicity, virality) of multi-modal inputs, and (2) a pragmatic logical inference model M_{EG} trained to generate factually grounded explanations of textual claim intent. The unified verification model is motivated by previous approaches that have shown transformer-based models can effectively learn task-specific and linguistic reasoning skills from unified pretraining for misinformation detection (Lee et al. [2021]). The inference model is motivated by our own work building on Frame Semantics. Using reinforcement learning, we can consider training M_{EG} to generate grounded explanations that maximize fact-checking performance. We can also condition explanation generation on user-specific cognitive models to oppose mis- and disinformation targeted at identity groups. These explanations could potentially aid human users as well as classifiers in identifying false or manipulative content.

Data contribution. In pursuit of distributional robustness, we would test the UMD system on an open source platform where real-world users could dynamically interact with models, verify a diverse pool of crowd-collected claims, and provide feedback on model performance. Such a setup would be ideal for gamification of model validation, empowering laypeople to be active participants in computational social science research and mitigating harmful content from online communities. It would also provide a rich database of information about detectors perform on temporally shifted data.

Systemic contribution.

Beyond providing new sources for data curation and democratization of model evaluation, the proposed UMD platform could provide a voice to online communities that are often rendered invisible. The impact of misinformation is shaped by many socio-cultural factors, including historical discrimination. For example, in the American Black community there are legitimate concerns about vaccination due to past and present racially motivated medical mistreatment (Martin et al. [2022]). Transforming research into and development of AI-powered systems into a community-driven process can take social nuances like this into account, leading to countermeasures that are personalized to affected audiences.

8.2 Conclusion

Given the challenges posed in the introduction, it seems that progress in safety and factuality of AI-powered systems requires a fundamental rethinking of our current learning paradigms. We posit that the most promising way forward relies upon more closely tying the training process with the intended user and end tasks. In future work we encourage addressing this issue with a focus on controllability and editability (Meng et al. [2022]), to think of the models underpinning these systems not as fixed points but as works-in-progress that should be adaptable to both distributional shifts in target data and socio-cultural shifts these can reflect.

Bibliography

Daron Acemoglu, Asuman E. Ozdaglar, and James Siderius. 2022. A model of online misinformation. *CEPR Discussion Paper*.

Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. The cringe loss: Learning what language not to model. *ArXiv*, abs/2211.05826.

E. Ageeva, M. Forcada, Francis M. Tyers, and Juan Antonio Pérez-Ortiz. 2015. Evaluating machine translation for assimilation via a gap-filling task. In *EAMT*.

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA. Association for Computing Machinery.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *ArXiv*, abs/2204.06031.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Peggy M. Andersen, Philip J. Hayes, Alison K. Huettner, Linda M. Schmandt, Irene B. Nirenburg, and Steven P. Weinstein. 1992. Automatic extraction of facts from press releases to generate news stories. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, page 170–177, USA. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Conference on Empirical Methods in Natural Language Processing*.

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 534–545, Doha, Qatar. Association for Computational Linguistics.

Ian A. Apperly. 2010. Mindreaders: The cognitive basis of "theory of mind". Psychology Press.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Conference on Empirical Methods in Natural Language Processing*.

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M.B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, 115:9216 – 9221.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, WWW '19, page 3286–3292, New York, NY, USA. Association for Computing Machinery.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Annual Meeting of the Association for Computational Linguistics*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *ICLR*.

Luke Breitfeller, Emily Ahn, Aldrian Obaja Muis, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Conference on Empirical Methods in Natural Language Processing*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NeurIPS*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. *In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Marilyn J. Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313.

Sapna Cheryan and Galen Von Bodenhausen. 2000. When positive stereotypes threaten intellectual performance: The psychological hazards of "model minority" status. *Psychological Science*, 11:399 – 402.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Annual Meeting of the Association for Computational Linguistics*.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *ArXiv*, abs/2006.00885.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

Alexander M. Czopp, Aaron C. Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10(4):451–463. PMID: 26177947.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *ICLR*.

Aida Mostafazadeh Davani, M. C. D'iaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv*, abs/2012.00614.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Conference on Empirical Methods in Natural Language Processing*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Annual Meeting of the Association for Computational Linguistics*.

Esin Durmus, He He, and Mona T. Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philip Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019a. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019b. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Don Fallis. 2014. A functional analysis of disinformation.

Lisa K. Fazio, Nadia M. Brashier, Brian K. Payne, and Elizabeth J. Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of experimental psychology. General*, 144 5:993–1002.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *AAAI*.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on twitter? *SIGMETRICS Perform. Eval. Rev.*, 44(1):179–192.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021a. Go figure: A meta evaluation of factuality in summarization. *Findings of ACL*.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2021b. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Annual Meeting of the Association for Computational Linguistics*.

Saadia Gabriel, Hamid Palangi, and Yejin Choi. 2022. Naturaladversaries: Can naturalistic adversaries be as effective as artificial adversaries? In *Conference on Empirical Methods in Natural Language Processing*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*.

Peter Glick and Susan T. Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70:491–512.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *In Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed

Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA. Association for Computing Machinery.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *ICLR*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20:818–829.

Ben Goodrich, Mohammad Saleh, Peter J. Liu, and Vinay Rao. 2019. Assessing the factual accuracy of text generation.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Conference on Automated Knowledge Base Construction*.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

H. Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Annual Meeting of the Association for Computational Linguistics*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Stuart Hall. 1973. Encoding and decoding in the television discourse. University of Birmingham.

William Hart, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa A Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135 4:555–88.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *ACL*.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *North American Chapter of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. 2019. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Annual Meeting of the Association for Computational Linguistics*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Annual Meeting of the Association for Computational Linguistics*.

Joe Hoover, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon, and Morteza Dehghani. 2019. Bound in hatred: The role of group-based morality in acts of hate.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean D. Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *NLP4COVID@EMNLP*.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. *2019 International Conference on Multimodal Interaction*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1587–1596. JMLR.org.

Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, page 43–58, New York, NY, USA. Association for Computing Machinery.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

David L Hudson, Jr. 2017. Hate speech online. `https://web.archive.org/web/20211115012316/https://www.freedomforuminstitute.org/first-amendment-center/topics/freedom-of-speech-2/internet-first-amendment/hate-speech-online/`. Accessed: 2021-11-14.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 2:1 – 23.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence*.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Muhammad Khalifa, Hady ElSahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. *ICLR*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Talat, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in nlp. *NAACL*.

Philipp Koehn. 2010. Statistical machine translation. Cambridge University Press.

Klaus Krippendorff. 1980. Content analysis: An introduction to its methodology.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Chinmay Kulkarni and Ed Chi. 2013. All the news that's fit to read: a study of social annotations for news reading. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Conference of the European Chapter of the Association for Computational Linguistics*.

J. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Erik G. Learned-Miller, Vicente Ordonez, Jamie Morgenstern, Joy Buolamwini, Sasha Costanza-Chock, Aaina Agarwal, Ben Hutchinson, Brant A. Cheikes, and David Evans. 2020. Facial recognition technologies in the wild: A call for a federal office.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *North American Chapter of the Association for Computational Linguistics*.

Res Lett. 2017. Fake news threatens a climate literate world. *Nature Communications*, 8.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Charles G. Lord, Lee D. Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2021. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. In *North American Chapter of the Association for Computational Linguistics*.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

Kimberlynn J. Martin, Annette L. Stanton, and Kerri L. Johnson. 2022. Current health care experiences, medical trust, and covid-19 vaccination intention and uptake in black and white americans. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Conference on Empirical Methods in Natural Language Processing*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020a. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020b. How decoding strategies affect the verifiability of generated text. *In Findings of the Association for Computational Linguistics: EMNLP 2020*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. *ArXiv*, abs/2009.06807.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004*

Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Clyde R. Miller. 1939. The techniques of propaganda. *How to Detect and Analyze Propaganda*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *ICLR*.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end nlp knowledge graph construction. *ICML*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Conference on Empirical Methods in Natural Language Processing*.

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *International Joint Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Conference on Computational Natural Language Learning*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Conference on Empirical Methods in Natural Language Processing*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175 – 220.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. 2019a. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *International Conference on the Web and Social Media*.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. 2019b. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *ArXiv*, abs/2102.04567.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. *ICWSM*.

OSTP. 2022. Blueprint for an ai bill of rights.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *EMNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res. 21, 1, Article 140*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Annual Meeting of the Association for Computational Linguistics*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Conference on Empirical Methods in Natural Language Processing*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Annual Meeting of the Association for Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35:529–558.

Zhiying Ren, Eugen Dimant, and Maurice E. Schweitzer. 2023. Beyond belief: How social engagement motives influence the spread of conspiracy theories. *Journal of Experimental Social Psychology*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Victoria L. Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news.

Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. 2022. On the robustness of offensive language classifiers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7424–7438, Dublin, Ireland. Association for Computational Linguistics.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. 2020. Do adversarially robust imagenet models transfer better? *NeurIPS*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI Conference on Artificial Intelligence*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019b. The risk of racial bias in hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022a. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022b. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *NAACL*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *North American Chapter of the Association for Computational Linguistics*.

Tal Schuster, R. Schuster, Darsh J. Shah, and Regina Barzilay. 2019. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, pages 1–18.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain James Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *ACL*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Conference on Empirical Methods in Natural Language Processing*.

Molly J. Simis, Haley Madden, Michael A. Cacciatore, and Sara K. Yeo. 2016. The lure of rationality: Why does the deficit model persist in science communication? *Public Understanding of Science*, 25(4):400–414. PMID: 27117768.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *International Conference on Language Resources and Evaluation*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Zeerak Talat, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp. *ArXiv*, abs/2101.11974.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *ArXiv*, abs/1910.08684.

Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Oken Hodas. 2017. Separating facts from fiction:

Linguistic models to classify suspicious and trusted news posts on twitter. In *Annual Meeting of the Association for Computational Linguistics*.

Soroush Vosoughi, Deb K. Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146 – 1151.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *ArXiv*, abs/2002.10957.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models

to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel C. Woolley. 2020. *Bots and Computational Propaganda: Automation for Communication and Control*, SSRC Anxieties of Democracy, page 89–110. Cambridge University Press.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. *Findings of EMNLP*.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *AAAI Conference on Artificial Intelligence*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *NeurIPS*.

Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018a. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 603–612, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ICLR*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018b. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. *ICLR*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Chenguang Zhu, William Fu-Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *North American Chapter of the Association for Computational Linguistics*.

Caleb Ziems, William B. Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *ArXiv*, abs/2305.03514.